DOCUMENT RESUME

ED 216 034 TM 820 191

AUTHOR	Haney, Walt; Scott, Laurie
TITLE	Talking with Children About Tests; A Pilot Study of Test Item Ambiguity. National Consortium on Testing Staff Circular No. 7.
INSTITUTION	Huron Inst., Cambridge, Mass.
SPONS AGENCY	National Consortium on Testing, Cambridge, Mass.
PUB DATE	Aug 80
NOTE	124p.
AVAILABLE FROM	The Huron Institute--National Consortium on Testing Project, 123 Mt. Auburn St., Cambridge, MA 02138 ($6.00).
EDRS PRICE	MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS	*Ambiguity; Grade 2; Grade 3; *Interviews; Item Analysis; Primary Education; *Standardized Tests; *Test Items; *Test Validity; *Young Children
IDENTIFIERS	California Achievement Tests; Comprehensive Tests of Basic Skills; Metropolitan Achievement Tests; Stanford Achievement Tests

ABSTRACT

A pilot study was conducted on test item ambiguity in standardized tests. A group of second and third grade children were selected and given a random sample of items from four standardized test series. Items were drawn from the reading, science and social studies subtests and put together in a composite reading comprehension, science, and social studies test. The children were interviewed individually about their perceptions of and answers to each item. The interviews were audiotape recorded and given analytical and holistic ratings. In both rating schemes, ambiguity in children's interactions with test items were of two types. The researchers attempted to distinguish three types of keyed answer ambiguity in both types of ratings; selection of a keyed answer by guessing, by mistake, or by other reasoning which did not reflect the skill measured. The limitations of the study were cited and explained. Ambiguity was seen as a real phenomenon, but substantially less of a problem than some critics indicated. The issues of ambiguity raised by this study refer more to validity than reliability. The authors suggest further research into children's perception and reasoning about tests and test items, as well as test interpretation and use. (DWH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

The Huron Institute
NATIONAL CONSORTIUM ON TESTING PROJECT
123 MOUNT AUBURN STREET
CAMBRIDGE, MASSACHUSETTS 02138
617-661-9285

ED216034

NATIONAL CONSORTIUM ON TESTING

STAFF CIRCULAR NO. 7

TALKING WITH CHILDREN ABOUT TESTS: A PILOT STUDY OF TEST ITEM AMBIGUITY

August 1980

Walt Haney
&
Laurie Scott
The Huron Institute
123 Mount Auburn Street
Cambridge, Ma. 02138

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL IN **MICROFICHE ONLY** HAS BEEN GRANTED BY

W. Haney

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

This is the seventh of a series of staff circulars distributed as part of The Huron Institute's staff work for the National Consortium on Testing. This circular is intended to spur discussion on the topic of children's perceptions and reasoning about tests, and to raise important issues worthy of additional attention, not to resolve them. It is produced primarily as a resource document for member organizations of the NCT. Although the circular has been prepared as part of staff work for the Consortium, it does not necessarily reflect the views of the Consortium nor of individual Consortium members.

© W. Haney for the NCT.

TM 820 191

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS iii

I. INTRODUCTION 1

II. TEST ITEM AMBIGUITY: A LONGSTANDING DEBATE 3

III. THE PILOT STUDY 13

Participants 14
Construction of Composite Tests 15
Administration of Composite Tests and Interviews 19
Standard Test Results 21

IV. ANALYSIS AND RESULTS 23

Analytical Ratings 24

- Reading Comprehension
- Science
- Social Studies

Analytical Rating Results 31

- Aggregate Results of Analytical Ratings
- Detailed Results of Analytical Ratings

Holistic Ratings 44
Holistic Ratings Results 47
Overall Findings 51

- Patterns of Results
- Consistency of Results
- Statistical Significance of Results

V. DISCUSSION: LIMITATIONS OF STUDY AND SOURCES OF AMBIGUITY 58

Interviews 58
Ratings 59

- Analytical Ratings
- Holistic Ratings

Sources of Ambiguity 66

- Rules of Test-Taking
- Unusual and Perceptive Interpretation
- Personal Experience and Egocentrism
- Concrete Sequential Reasoning
- Visual Interpretation
- Item Format

The Ambiguity of Ambiguity 89

VI. SUMMARY AND CONCLUSIONS 93

Background 93
The Pilot Study 93
Analysis 94
Discussion 98
Conclusions 102

REFERENCES 113

ACKNOWLEDGEMENTS

Though the study reported in this paper was fairly small, the help of numerous people was essential in carrying it out. Thanks go first to Banesh Hoffmann whose gentle but insistent prodding led us to undertake the study. The helpfulness of Susan Kluver and Ken Schmidt was essential to our carrying out the study and is much appreciated. Thanks go too to Emily Cahan, Mark McQuillan, Elice Formann and Mieko Kamoi for their help in conducting the study. For their useful comments and criticisms, we express thanks to Anne Anastasi, Mike Garet, David Cohen, George Madaus, Vito Perrone, Michael Beck, Elinor Woods, and Anne Smith. Special appreciation is owed to Norma Robbins who cheerfully and efficiently worked through at least a half-dozen versions of this manuscript. Most of all, however, we wish to express thanks to eleven children, whom we can only call A - K for present purposes.

None of these people are, of course, responsible for any errors or ambiguities which remain in the paper.

Walt Haney

Laurie Scott

I. INTRODUCTION

> No sense, no logic, no reality. That is the impression you get from reading through test after standardized test. At first you think there must be some mistake, some one or two test makers who do a particularly poor job. And some tests *are* truly terrible. But *all* of them I have read are at least bad.
>
> Edwin Taylor
> "The Looking-Glass World of Testing"
> 1977a p.11.

> Articles in popular and even some semi-professional media similarly criticize individual items on familiar grounds: no right answer, more than one right answer, unfairness to unconventional types of thinkers, ambiguity, dependence on factors other than those supposedly being measured, etc. Test makers respond as you would expect; instruments, like any human products, are not perfect; given the enormous number of test items in all the published tests, some poor items find their way into such tests despite maintenance of a generally high level of item quality through sophisticated editorial review and item analysis activities. Unless it is demonstrated that such items make up a significant part of a test, their existence does not vitiate its usefulness.
>
> Roger T. Lennon
> "Perspective on Intelligence Testing"
> 1978, p.6.

Recent debate about standardized testing covers a wide range of educational, political, scientific and economic issues. But one recurring theme in the debate concerns the quality of tests and test items. Critics, such as Edwin Taylor, charge that standardized test items often are ambiguous, misleading and nonsensical. Test publishers, such as Roger Lennon, respond by saying that perhaps a few poor items do find their way into tests, but their number appears to be few and so it is unfair and inaccurate to characterize all standardized tests as ambiguous and misleading by pinpointing attention on just a few such items.

Empirical evidence can be brought to bear on this dispute. This proposition was the idea behind the pilot study reported in this paper. Specifically the purpose of the study was to gather empirical evidence on the prevalence of ambiguity in commonly used standardized test items. The idea was simple. First, we drew a random sample of test items from several of the most commonly used standardized tests. Second, we administered the test items to a small group of children. Third, we talked with each child about how he or she answered each item. Fourth, we analyzed children's accounts of how they answered items in an attempt to estimate the frequency with which they answered items for reasons which were irrelevant to the skill or attribute ostensibly being measured by the item.

Though the idea for the pilot study was quite simple, carrying it out was not, as we explain below. First, however, we should make clear that the idea of talking with children about how they make sense of standardized test items is not new. In fact it goes back at least half a century. In Section II, we review previous research along these lines and point out exactly what it is that the pilot study was intended to contribute to past research. In Section III, we describe how the study was designed and carried out. Section IV describes two different approaches to analyzing children's accounts of how they answered test items and presents the results of the study. Section V provides a discussion of the pilot study findings, including a review of the problems and difficulties encountered in the study. Some of these will also be discussed briefly in Section III since they are relevant to interpretation of results presented in Section IV. Section VI provides a summary and our conclusions from the pilot study.

Before setting out the substance of these sections, we should make clear that the work reported here represents only a pilot study. It was designed simply to bring some initial evidence to bear on the question of the prevalence of test item ambiguity. Nevertheless, even as a tentative inquiry the study has yielded some provocative findings. These findings will, we hope, both shed some light on the interpretation of test results and serve to spur additional research along similar lines.

II. TEST ITEM AMBIGUITY: A LONGSTANDING DEBATE

Concern over the quality of standardized test items and how they are interpreted goes back at least 50 years. Test item quality was one of the points of contention in the famous series of exchanges on IQ testing between Walter Lippmann and Lewis Terman in the 1920s (see Block and Dworkin, 1976). Also around 1920, Jean Piaget began his long career in psychology by talking with Parisian school children about how they made sense of test questions:

> Instead of administering the tests in a standardized way, he interviewed the children at length. The children did not complain. They found this "M'sier" amusing. With him, one was not bored, at all. Instead of simply noting the responses given by the children to the test items, the young Swiss biologist was interested in the how and the why of the answers. What had been at the outset nothing but a boring and annoying test situation became a real dialogue with suggestions and countersuggestions, an argument developed, a deepening of the child's thought, a new method of interrogating children was born. It leads the child to show how he formulates and solves a problem, how he thinks.
>
> (Gruber & Voneche, 1977, p.53)

M'sier Piaget, of course, used this approach not to study the nature of tests, but instead to investigate the nature of children's thinking.

Talking with children about why they answer test questions in particular ways has been advocated by several observers as a way of evaluating test quality. In The Tyranny of Testing (1962), Banesh Hoffmann suggested an experiment in which students would be engaged in analyzing test quality to seek out ambiguities and errors in test items (pp. 181-2). Anne Anastasi has noted that test validation may include "analyses of types of errors commonly made on a test and observation of the work methods employed by examinees. The latter could be done by testing students individually with instructions to 'think aloud' while solving each problem" (Anastasi, 1976, p. 138).

Taylor has recommended that "test developers must sit down with children individually, watch them take the test, and talk with them afterward about which questions were clear and important and which were confusing or demeaning" (Taylor, 1977, p.14).

Systematic studies along these lines appear to be quite rare. Far more common is for adult observers themselves to review tests in order to seek out items that are ambiguous, misleading or otherwise inadequate. Indeed this practice is common among critics of standardized testing (see Hoffmann, 1962; Patton, 1975; Taylor, 1977; Schwartz, 1977; and others in Houts, 1977). Taylor, for example, cites the following items from published science tests as examples of ambiguous and misleading questions:

> Ocean waves are caused by the
>
> earth's gravity
> moon's gravity
> earth's motion
> wind's patterns
>
> Many kinds of plants are not able to live
> on the desert because of the
>
> high temperature
> low rainfall
> bright sunshine
> poor soil
>
> (Taylor, 1977b, pp. 303-305)

The practice of having adults review tests intended for children is by no means exclusive to standardized testing critics, however. Publishers of standardized tests typically employ a variety of adults to review items in the process of test development. Commonly this includes reviews by subject matter specialists, measurement experts and, especially in recent years, minority group representatives.

After candidate items are reviewed by a variety of experts, developers typically pretest items by having a sample of children take a tryout test. Candidate items are then chosen for inclusion in the final version of the test on the basis of item tryout statistics such as item difficulties (often called the p-value or percentage of test-takers answering the item correctly) or item discrimination indices (measures of the extent to which item results differentiate between test-takers who score high and low on the total test; see Anastasi, 1976, pp. 206-215, or Henrysson, 1971, for details on how such indices are calculated). Defenders of standardized tests often claim that test critics ignore such procedures. In his critique of Hoffmann's book The Tyranny of Testing, Dunnelle (1964), for example, charged that:

> . . . Hoffmann completely ignores the fact that statistical validation of test items is most often an effective means of discovering poor and ambiguous items. Nowhere in his book does Hoffmann mention that item analysis is primarily a means of identifying poorly keyed and ambiguous items (p.66)

Despite this disclaimer, these procedures do not necessarily get at the issue of test or item quality from the perspective of individual test-takers which was the point of Hoffmann's critique. This distinction is important because what it is that a test or item measures--that is, its content validity--depends not on what adult experts or critics think it measures nor on what item statistics suggest about the item but rather on how individual test-takers perceive and react to the test or item. As Anastasi points out, "Content validity depends on the relevance of the individual's test responses to the behavior area under consideration, rather than on the apparent relevance of item content. Mere inspection of the test may fail to reveal the processes actually used by examinees in taking the test" (Anastasi, 1976, p. 135).

To delve into what it is that a test or test item measures for particular test-takers requires some kind of observation or communication with them on an individual basis. This is precisely the sort of interchange which takes place between test administrator and test-taker in many sorts of individualized testing, and textbooks on individually administered tests suggest a variety of approaches to dealing with unusual and ambiguous answers to individually administered test questions. Matarazzo (1972), for example, in his text Wechsler's Measurement and Appraisal of Adult Intelligence, offers the following advice in a discussion of the scoring of the Picture Arrangement Test of the Wechsler Adult Intelligence Scales:

> The arrangements given in the manuals cover pretty well most of the rational orders which the individual series permits. Occasionally, however, a subject does produce a different one for which he is able to give a convincing explanation, but for which no credit is allowed in the manual. In most instances, it will be found that disallowing the subject's response does not materially influence his total score on the test, but provision is made for the examiner to credit the subject with a reasonable additional score in special cases. More interesting than the question of credits allowed, in such cases, is the explanation which the subject may give for his unusual arrangement. Consistently bizarre explanations are suggestive of some peculiar mental orientation or even psychotic process.
>
> (Matarazzo, 1972, p. 210)

Later in a discussion of extraneous factors which can influence performance on intelligence tests, he argues:

> Emotionality, anxiety, motivation, introversion-extraversion, etc., can influence test scores,

> but only seldom do they influence performance to such a degree as to invalidate the test findings as a whole. More important..., however is the fact that the impacts of these personality factors, far from being sources of error, must be looked upon as significant aspects of the individual's global intellectual capacity.
>
> (Matarazzo, 1972, p. 484 emphasis in original)

Though unusual and ambiguous interpretation of individually administered test items has thus been clearly recognized, there appears to have been little systematic research on this point from the perspective of individual test-takers, particularly with respect to group-administered tests. In the remainder of this section we describe briefly past work along these lines and identify what the pilot study was intended to contribute to it. In the concluding section, we will refer also to additional relevant research, some of which came to our attention only after our investigation was begun.

In its early years, the National Association of Educational Assessment (NAEP) undertook an unusually wide range of investigations in the course of item or, to use NAEP's term, exercise development. One of them involved talking with school children after they had taken the NAEP exercises:

> A small number of examinees, judged to be the least able in each class, were interviewed carefully after each group administration. They were asked about the exercises and any problems they encountered. The results of the study are based primarily upon interviewer, teacher and student comments.
>
> The major conclusion was that the exercises, as they existed in the spring of 1967, were basically usable but needed revision. . . .
>
> The major problems with the exercises were directions that were too complex or involved and use of vocabulary that was so difficult that low achieving examinees had trouble understanding the task or the question. The post-test interviews were particu-

> larly productive in identifying specific words, phrases and directions that were incomprehensible to many students.
>
> (Finley & Berdie, 1970, pp. 61-62)

A second set of investigations involving talking with children about how they made sense of test items was reported to Mehan (1973), MacKay (1973), Mehan and Wood (1973), and Cicourel et al. (1974). In related studies, these investigators studied how children interpreted and answered a variety of standardized test questions:

> After standardized reading, intelligence, and language development tests were given to the first grade children in both schools, a small group of children were asked to explain their answers and to provide the reasons why they answered as they did. The children's accounts were then compared to the official versions provided by test protocol. As in the classroom lesson situation, the children were not always attending to testing material in the way expected by the test protocol. Often this differential perception led the child to a wrong answer, which in turn led to a low test score.
>
> (Cicourel, et al., 1974, p.5)

In summary, these investigators challenged

> the conclusion usually drawn from incorrect test scores: that the child being examined lacked particular abilities. Instead they suggest that it is necessary to examine the structure of the child's accounting practices and reasoning processes in order to draw valid inferences about his competence.
>
> (Cicourel, et al., 1974, p.5)

Roth's (1974) study employed the individually administered Peabody Picture Vocabulary Test. After analyzing a sample of children's accounts of how they made sense of questions on this test, Roth observed that "the Peabody Test is insufficiently verbal because it limits the child's verbal performance to matching picture and word. When the child is encouraged to verbalize about the pictures and words, we find that he knows much more than the test score indicates" (p. 203, emphasis in original). Roth reported that this is true for both black and white, and low-scoring as well as high scoring children.

In a comparison study MacKay (1974) analyzed how first-grade children arrived at answers to the Cooperative Reading Test (Form 12):

> After the children had taken the test I went over it with them, asking them to read the stimulus word or sentence and to tell me why they had chosen the answer they marked. The interviews were audio tape recorded and later transcribed.
>
> (MacKay, 1974, p.220).

Though he does not report results of his study thoroughly, MacKay does provide several examples of what he calls ambiguous items:

> Item 18 is one example of what I am referring to [ambiguity of test items]. The stimulus sentence is "The bird built his own house" followed by pictures of (1) a twig; (2) a birdhouse; and (3) a nest. According to the national norms in the Manual the distribution of answers was: (1) 5%; (2) 59%; and (3) 35%. Number 3 (the nest) was the correct answer. If the relevant skill being measured is understood to be "Identifying an illustrative instance" then the choice of 2 or 3 seems warrantable. Indeed most children seem to have thought Number 2 to be most reasonable. The teacher commented on this in an interview after the test was administered:
>
> > "I saw somebody read the whole thing to himself 'The bird build his own' and they checked the bird house because you build a bird house and bird built ... his own house -- the word house throws them so 'bird built his own house,' well in fairy stories quite often a rabbit builds his own house. There's no reasons why in make-believe. . . a bird couldn't [build] his own house. Monkeys do it. We just read a story where monkeys didn't want to build their house. To them house is misleading. If it said the bird built its own nest they would have checked that."

> Perhaps birds do build birdhouses in Walt Disney cartoons and that is the frame of reference the child uses. The test constructor's power is to enforce narrow normative conceptions.
>
> (MacKay, 1974, p. 244, teacher's account edited from original)

In conclusion MacKay suggests

> the test fails within its own frame of reference because of lack of clear instructions, lack of standardization, lack of accurate time keeping and most importantly ambiguity in many of the items. (p. 245)

Mehan (1978) reports on these and other similar investigations of how children make sense of test items. Again results are not reported comprehensively. "Because of space limitations, only a few examples of the disparity between adult and child perspectives will be presented here" (p. 50). After presenting two such examples, Mehan observes:

> These answers did not result from a lack of knowledge; rather, they resulted from a substantively different interpretation of test materials. Students who answered test questions incorrectly were often performing the very cognitive operation being tested by the questions. Contrary to prevailing educational-testing theory, which suggests that incorrect answers may result from a lack of knowledge, these investigations of children's schemes of interpretation suggest that incorrect answers may result from a discrepancy between adult and student views of the world. (p. 51)

In a booklet entitled "Reading Tests: Do They Help or Hurt Your Children?" Cook and Meir (n.d.) report on items from the Metropolitan Achievement Test used with children aged 7 to 9 years. They note numerous instances in which items can be confusing or misleading to young children. Though they do not say how the examples cited were selected, they do suggest that "the test questions in this booklet are typical of those found on all standardized reading tests--not just the MATs." (See also Meier, 1972)

In sum, critics of standardized tests (Hoffmann, 1962; MacKay, 1973, 1974; Cook and Meier, n.d.; Mehan, 1973, 1978) have charged that standardized test items frequently are misleading and ambiguous and do not necessarily measure what they purport to measure -- that is, that they lack content validity.

Defenders of standardized tests (e.g., Dunnelle, 1964; Ebel, 1973; Lennon, 1978; Jensen, 1980) have responded to these charges in three ways:* by arguing that:

1. Such critisms ignore the fact that expert review, item tryout and test validation procedures help to ensure that poor and ambiguous test items do not find their way into published tests.
2. What may appear ambiguous to an adult may not seem so to a child because adult perceptions of test items do not necessarily reflect how children perceive and experience those items.
3. Even if a few test items are ambiguous or misleading there is no systematic evidence to indicate that they constitute a significant portion of items on any one test, or on standardized tests generally.

The first response seems plausible, yet the fact that critics have identified ambiguous or otherwise poor items among those included on published and widely used standardized tests indicates that such procedures are not infallible means of screening out such items. The second point is relevant also. Adults' perceptions of test items do not however necessarily pertain to how items are experienced by those with whom they are intended to be used. But the converse of the second point seems equally relevant--what seem ambiguous or misleading from the child's perspective may not seem so from the adult's viewpoint. Very few published studies appear to have addressed the question of ambiguity from the perspective of the test-taker. Among the best of these studies are the related work of Mehan (1973), MacKay (1973, 1974), and Roth (1974). Their studies with first graders in California have demonstrated that children may choose "correct" or intended response alternatives without actually applying the skill or knowledge ostensibly being measured by a test item and conversely that a wrong or unintended response alternative may be selected even though the child correctly applies the skill or knowledge ostensibly being measured by the item.

---

* Actually, defenders of standardized tests have responded in numerous other ways to the test critics. Other responses deal with such things as social philosophy, and educational theory (e.g., see Dunnelle, 1964; Ebel,1973). The three points listed above seem to us to represent only the direct substantive responses of the defenders to the critics' charges of ambiguity.

What is unclear from these studies, however, is the third point in the test defender's brief for the defense; namely the extent to which these findings are generalizable. Though some observers, after recounting several examples of ambiguous or misleading items, have drawn rather sweeping conclusions about all items on a test or even items on standardized tests generally, the prevalence of items which are misleading or ambiguous, from the point of view of the test-taker, has not been systematically investigated. The pilot study reported in this paper was designed to provide empirical evidence on precisely this point.

III. THE PILOT STUDY

The pilot study was conducted with second- and third-grade children. using items drawn from four of the most commonly used standardized test series. In this section we describe: 1) how children were selected for participation in the study, 2) how test items were selected and put together into composite tests, 3) how these tests were administered, 4) how children were interviewed about the tests, and 5) how children scored on the test.

3.1 Participants

Eleven children participated in the pilot study. All were in the same combined second and third grade class. The six girls and five boys ranged in age from seven years eight months to ten years five months. Four were in the second grade and seven in the third.

These children were not systematically selected from any broader group of children. Initially, we met with a group of parents from the school in which the study was to be conducted and solicited their interest in the pilot study. We discussed with them some of the common criticisms of standardized testing, as well as some of the reasons why standardized tests are widely used. We also described our plans for carrying out the pilot study. Parents who approved of the study plan were asked to allow their children to participate. Only those students whose parents provided written approval were included in the study.

The sample included both black and white children and appeared to include a moderate range in socioeconomic background. No data on race or SES were included in our analyses however. Since the study was designed as a pilot inquiry, our intent was simply to find a group of fairly typical school children willing and able to participate in the study. Identities of the participants must, of course, remain confidential. For purposes of analysis and reporting they are referred to simply as students A-K.

While the eleven students participating in the pilot study seemed to represent a fairly typical sample of lower to upper middle class school children, one respect in which they may have been somewhat unusual should be noted. All came from a relatively open classroom in which children were encouraged to speak freely and to question and explore their educational interests. This educational background may have made it easier for the children to discuss the test items candidly with us and may have affected the way they perceived and reacted to standardized tests. Nevertheless, all of the children were reported to have taken standardized tests on previous occasions.

3.2 Construction of Composite Tests

Items used in the pilot study were sampled from four widely used standardized achievement test series, namely the Metropolitan Achievement Test (MAT); the Comprehensive Test of Basic Skills (CTBS); the California Achievement Test (CAT); and the Stanford Achievement Test (SAT). According to three independent surveys,these are four of the most widely used standardized achievement test series in the United States. These three surveys are (1) a survey of test use in the nation's 100 largest school districts (Dimengo, 1978); (2) an analysis of instruments used in federally-sponsored research on children and adolescents (Heyneman and Mintz, 1976); and (3) a review of state Title I annual evaluation reports (Gamel et al., 1975).

Relevant test levels were first identified from each of these test series (see Table 1). We attempted to select test levels appropriate for the pilot study sample; that is, levels which were appropriate according to the publishers' designations, for roughly middle of second

TABLE 1. STANDARDIZED ACHIEVEMENT TESTS USED

Test Series	Publisher	Copyright Dates	Level	Intended Grade Range	Form	Subtests	No. of Items
Metropolitan Achievement Test (MAT)	Harcourt Brace Jovanovich, Inc.	1977-1978	Primary 2	2.5-3.4	JS	Reading Science Social Studies	55 40 40
Comprehensive Tests of Basic Skills (CTBS)	CTB/McGraw-Hill	1973	C	1.6-2.9	S	Reading Comprehension: Passages Science Social Studies	 18 30 30
California Achievement Test (CAT)	CTB/McGraw-Hill	1977-1978	12	1.5-2.9	C	Reading Comprehension	20
Stanford Achievement Test (SAT)	Harcourt Brace Jovanovich, Inc.	1972-1975	Primary II	2.5-3.4	A	Reading Part B Science Social Studies	48 27 27

grade through end of third grade. In two cases, however, no test level spanned this range, so levels for a slightly lower grade range were selected. This seemed preferable to choosing a higher level which might have proven especially difficult for some of the second graders in our sample.

Three subtest areas were treated in the pilot study: reading comprehension, science and social studies. Each test series at the second-third grade level includes several other subtests, dealing, for example, with spelling, math computations, and reading vocabulary. The subtests of reading comprehension, science, and social studies were chosen for inclusion in the pilot study for two reasons. First, for practical reasons we had to limit the areas to be covered. Second, it was assumed that if ambiguity in test items was to show up anywhere it would be in the more complex subject matter areas of reading comprehension, science and social studies, rather than in the more concrete subject areas of vocabulary and math computations.

Reading comprehension is treated slightly differently in the four test series included in the study. For the sake of uniformity, we therefore defined comprehension for the purposes of this study to refer strictly to paragraph or passage comprehension, not to sentence or word comprehension. Science and social studies subtests are included at the second-third grade levels in only three of the four test series treated in the pilot study, namely the MAT, CTBS, and the SAT. The total pool of items defined by the relevant test levels was:

- Reading Comprehension	141
- Science	97
- Social Studies	97

Rather than administering all of these items, we selected a random sample of them. Using a formula for determining sample sizes (Krejcie & Morgan, 1970), we determined the number of items which would be necessary to attain a 0.10 accuracy of estimation with a confidence level of .80 (29 for the social studies and science item pools and 32 for the reading items). A random number table was then used to select science and social studies items from each test series so that the proportion selected from each corresponded to the proportion of each series' items in the total pool.

Reading comprehension items were selected in a slightly different fashion. Reading items are not independent of one another in the tests considered. Instead several items (typically four) are based on a single reading passage. If we had selected reading items on a strictly random basis, this would have required children to do more reading per item than is generally required in the actual tests. This likely would have made items more difficult than normal since in each of the test series studied, children are required to read only one passage in order to answer several items. Therefore we employed a stratified random sampling approach in selecting reading comprehension items. Entire reading passages were randomly chosen from each of the test series. As many complete sets of items (i.e., the reading passage plus all attendant questions) were selected as necessary to provide the requisite number of items. In some cases more than the required number of reading items were included. We could have selected only the exact number of items from each series, but this would have required splitting up some item sets (e.g., taking only two of four items pertaining to a particular reading passage).

Since this might have affected the difficulty of these items, we instead decided to take entire item sets, even though the total number of reading comprehension items thus selected was somewhat greater (37) then originally intended (32). We should note that the use of this stratified sampling procedure with the reading comprehension items has the effect of somewhat reducing the statistical power of the inferences which can be drawn with respect to the total pool of reading items. Nevertheless, this was deemed preferable to either making the test unreasonably long or selecting reading items individually rather than in sets as they normally occur in the tests.

Once items, together with relevant illustrations and reading passages, were selected, they were assembled into composite reading, science, and social studies tests. Clear xerox copies of individual items were taped together; and xerox copies of these "galleys" were then stapled together to form test booklets for each child. Although in constructing these composite test booklets we attempted to make reproductions as clear as possible, it should be noted that in some instances xeroxed illustrations in the composite tests were not as clear as were the originals in the actual test series. These booklets also included sample questions drawn verbatim from the original test series.

3.3 Administration of Composite Tests and Interviews

The three composite subtests were administered under standard testing conditions on three separate days. Each subtest required about half an hour to administer. Following group administration, each child was interviewed individually about each question; with interviews audio tape-recorded for later transcription. The two interviewers both had several years teaching experience and advanced degrees in education, but no special training in interviewing young children. Following a brief conversation intended to put children at ease, each child was given the subtest he or she had just completed. For the social studies and science subtests, each child was asked to describe the alternative answer choices pictured and then, after the interviewer repeated the test question, to explain his or her reasons for selecting a particular response. For the reading subtest each child was asked to read the question and alternative answer choices aloud (in only a small sample of cases were children asked to read the reading passages out loud) and then to explain reasons for answer selection. Duration of the interviews varied but generally lasted from 25 to 45 minutes for each child, for each subtest.

Absenteeism reduced the number of students participating for two subtests. Nine children took the social studies subtest; ten children took, the reading comprehension subtest, and all eleven took the science subtest.

All three subtests were administered after lunch. Most of the interviews took place either in the afternoon immediately after the testing or the following morning. The science interviews were begun on a Thursday afternoon and concluded on Friday morning. The students were noticeably more alert and attentive in the morning than in the afternoon, thus raising the question of the effect that time of day might have on both test and interview responses.

It should be noted that children's reactions to the interview process varied considerably. While most children seemed interested and attentive during interviews, a few grew obviously restless about having to respond to much the same questions for each of thirty test items. Also, in one case, an interview had to be terminated without completing discussion of all items simply because the school day came to a close and children had to depart to catch school buses.

3.4 Standard Test Results

Before describing how interviews with children were analyzed, let us simply summarize the overall performance of the pilot study sample on each of the composite tests. Results of the standard administration of the composite tests are shown in Table 2. The sample of children to whom the test was administered selected the keyed answer (that is the one designated by the publisher as the correct answer), for 72%, 67% and 71% of the items for the reading comprehension, science and social studies tests respectively. Since the composite tests were constructed from items randomly drawn from complete tests, it is impossible to compare this sample performance directly with test norms. However, since the average item difficulty reported by publishers for the items included in the composite tests was .64, .63 and .70 for reading comprehension, science, and social studies items respectively, the performance of the pilot study sample seems to correspond at least roughly with typical performance of the national norms samples for these test series.*

One other point should be noted from Table 2. On each of the composite tests a fair number of items were skipped by children (ranging from 31 or 3% of items on the reading comprehension subtest to 7 or 0.9% on the social studies subtest). In many cases students skipped items because, as they

---

* Average difficulties based on spring second grade norms sample or fall third grade where only these data were available. Note that while average number of items correct for the pilot sample is slightly higher then average item difficulties, seven of eleven students were third graders.

TABLE 2: STANDARD RESULTS ON COMPOSITE TESTS

	Composite Test		
	Reading Comprehension[1]	Science	Social Studies[2]
Sample Characteristics			
No. of Students	10	11	8
No. of Items	37	29	29
Total No. of Item Responses	358	319	232
Test Performance			
Keyed Answer Chosen			
Total	258	214	164
Mean/Child	25.8	19.4	20.6
%of Items	72	67	71
Unkeyed Answer Chosen			
Total	69	85	61
Mean/Child	6.9	7.7	7.6
% of Items	19	27	26
Item Skipped			
Total	31	20	7
Mean/Child	3.1	1.8	0.9
% of Items	9	6	3

---

1. Reading Comprehension test performance is reported on a total of only 358 since twelve items for one child were not included in the interview process.

2. Excluded from this summary of results is the social studies test for one student whose interview on this test was used in the training of raters for the analytical rating of interviews with children.

later explained they "couldn't get it" or "couldn't read it." In two instances, children skipped items because of clerical mistakes: one simply forgot to mark an answer and one girl inadvertently skipped a whole page of five items, all of which in the subsequent interview she was able to answer correctly.

IV. ANALYSIS AND RESULTS

Ambiguity refers to uncertainty or doubtfulness of meaning. For the purposes of our investigation, ambiguity in children's perception and reasoning about test items was defined in terms of contrasts between the standard scoring of items reported above, and our interviews with children. Two types of contrasts served to define ambiguity: 1) when a child marked the "right" answer, but the interview indicated he or she had not applied the general skill ostensibly measured by the item, or 2) when a child marked an "incorrect" answer, but the interview indicated that he or she had in fact applied the skill ostensibly measured. We will explore shortly exactly how these conditions were identified in terms of our analyses. Nevertheless, we wish to make clear from the start that this general approach includes certain phenomena that often might not be considered signs of ambiguity and excludes others that might reasonably be considered to be signs of just that condition. Nevertheless, this seems to us a reasonable place to start. Later, in Sections V and VI we will discuss alternative approaches to defining and identifying ambiguity in test items.

Tape recordings of interviews with individual children were transcribed and verified by checking them against the tape recordings. These transcriptions of interviews served as primary source materials for all subsequent analyses. Two types of analyses were performed on the transcriptions of interviews with children. The first was an analytical rating of each child's interview transcript, item by item in the order in which items appeared on each composite test. The second--what we call a holistic rating--was based on sets of item interviews, comprised of the test item and item statistics, together with an account of how all children interviewed described their perception and means of answering that particular item. In this section, we first describe the analytical rating and results, then the holistic rating and results. The concluding part of this section describes overall findings from both rating systems, in terms of how results from the two approaches compare.

4.1 Analytical Ratings

The analytical ratings consisted essentially of having two individuals (who had not been involved either in administration of the composite tests or in interviewing children about the tests) independently rate each account of how a child answered each item, using an analytical rating scheme. The two raters were both professional staff members of the Huron Institute, one trained in developmental psychology and the other in language and reading. In a moment we will describe the rating scheme, how it was applied, and how ambiguity was defined in terms of analytical ratings. First, however, let us simply describe the issues addressed in the analytical rating system.

The analytical rating scheme deals with aspects of children's thinking about test items which have been suggested to affect children's test performance. Specifically the six aspects of the rating scheme dealt with: (1) the answer alternative selected; (2) the child's perception of the item stimulus; (3) the child's perception of the response alternatives; (4) the process by which the child appeared to reach an answer to the question; (5) if the child appeared to be reasoning in answering the item question, the apparent validity of the logic used in light of the attribute ostensibly measured by the item; and (6) whether or not the child cited information or applied knowledge extraneous to the item or what it ostensibly measured.* These were elaborated as follows:

> Answer Alternative Selected. This was the most straightforward aspect of the analytical rating scheme; namely whether the child (1) marked the answer alternative which the test publisher's answer key indicated as correct (the keyed answer, (K); (2) marked a different answer choice, indicated by the publisher to be incorrect (an unkeyed answer, UK), or (3) no answer alternative was marked (skipped, S). Note that hereafter, we refer to "keyed" and "unkeyed" answer alternatives to indicate publishers' designations of correct and incorrect answers.
>
> Perceptions of Item Stimulus. This refers to how the child perceives what is depicted, presented and/or asked in the item. Raters were instructed to judge whether the child perceived the stimulus as apparently intended (or if unintended, that is where the perception was not one apparently intended by the item author, and was not directly related to the skill ostensibly measured by the item). These categories are abbreviated as IPE and UPS

---

* We should note that our rating scheme originally included a seventh issue, namely whether or not the child's account indicated that an item required any knowledge or behavior other than that ostensibly measured by the item or provided in the item stimulus. We included this because some test critics have charged that some items implicitly require such knowledge and that such requirements can affect test-takers' performance in a manner irrelevant to the attribute ostensibly being measured. In no cases, however, did a rater indicate that a child's response revealed that an item assumed such extraneous knowledge. This is a substantive finding but since there was no variation with respect to this issue in the analytical ratings, we will not discuss it further in this section. Nevertheless, we will come back to this point in Section V, since some items clearly do require extraneous knowledge, even if children's responses do not reveal this requirement.

in this paper. What we were trying to get at here was whether or not the child perceived the item stimulus in unintended fashion which was unrelated to the attribute ostensibly measured by the item. This rating category was not applied with reading items since we decided that there would be no way in which children's reading aloud of reading passages would reveal unintended perceptions unrelated to the skill of reading comprehension.

Perception of Response Alternatives. Children's perceptions of response alternatives were rated in similar manner as intended or unintended (abbreviated as IPR or UPR). Thus on a reading test item if a child could not read one response alternative, his perception would be rated intended, because the perception of the response alternative related directly to the skill ostensibly measured. However, if a child perceived a picture in unintended fashion on a science item this would be rated UPR, since visual perception was not part of what the item ostensibly measured.

Process of Reaching Answer. Under this category, raters were asked to judge what process the child used to reach an answer to a question. Specifically raters were asked to indicate whether or not the child evidenced some direct logic or reasoning (R), derived the answer indirectly via process of elimination of alternative answers (E); guessed an answer or at least said he or she guessed (G); or when explaining his or her answer spontaneously recognized that he or she had made a mistake when taking the test under standard conditions and corrected him or herself without assistance from the adult interviewer (M).

Reasoning. If the child was judged to have applied some sort of direct logic or reasoning in reaching an answer, the raters were further asked to judge whether the reasoning was valid (V) or invalid (Inv.) in light of the child's perceptions of the item stimulus and response alternatives. Note that it was possible for a child's reasoning to be rated V or Inv, regardless of whether or not he or she selected the keyed or unkeyed response alternative.

Extraneous Knowledge. If the child's reasoning was rated valid, the raters were further asked to judge whether the child's logic exhibited knowledge or behavior extraneous to the attribute ostensibly measured (E) or whether the child's logic was directly related to the attribute ostensibly measured (NE).

Ratings on several of these issues obviously depend on what it is that the test items ostensibly measure. On this point raters were instructed to rely directly upon publishers statements of what it is that these reading comprehension, social studies and science items measure. Publishers' general descriptions of what each of the three types of test items measure formed the basis for raters' judgments about what items ostensibly measured. These descriptions may be summarized as follows for each of the three types of items:

Reading Comprehension. Reading comprehension subtests (specifically paragraph or passage comprehension), according to manuals of the four test series employed, test two components of reading: literal comprehension and implicit, inferential or critical comprehension. The SAT Comprehension subtest employs completion items in which students are asked to choose one of four words to complete a sentence. The other three series use complete reading passages followed by a set of multiple choice items pertaining to the passage.

Science. The MAT Science items are intended to measure knowledge, comprehension, inquiry skills, and critical analysis in physical science, earth and space science, and life science (Prescott et al., 1978). The SAT is designed to test concepts of matter and energy, change in the physical environment, form and function of living things, environmental interaction, and basic skills of science (Madden, et al., 1973). The CTBS items are intended to "assess the student's ability to investigate problems in science and, to a lesser degree, to recall scientific facts or concepts" (CTB/McGraw-Hill, 1976, p. 5). Science items from all three series are intended to be read aloud by the examiner, and response choices are in the form of illustrations.

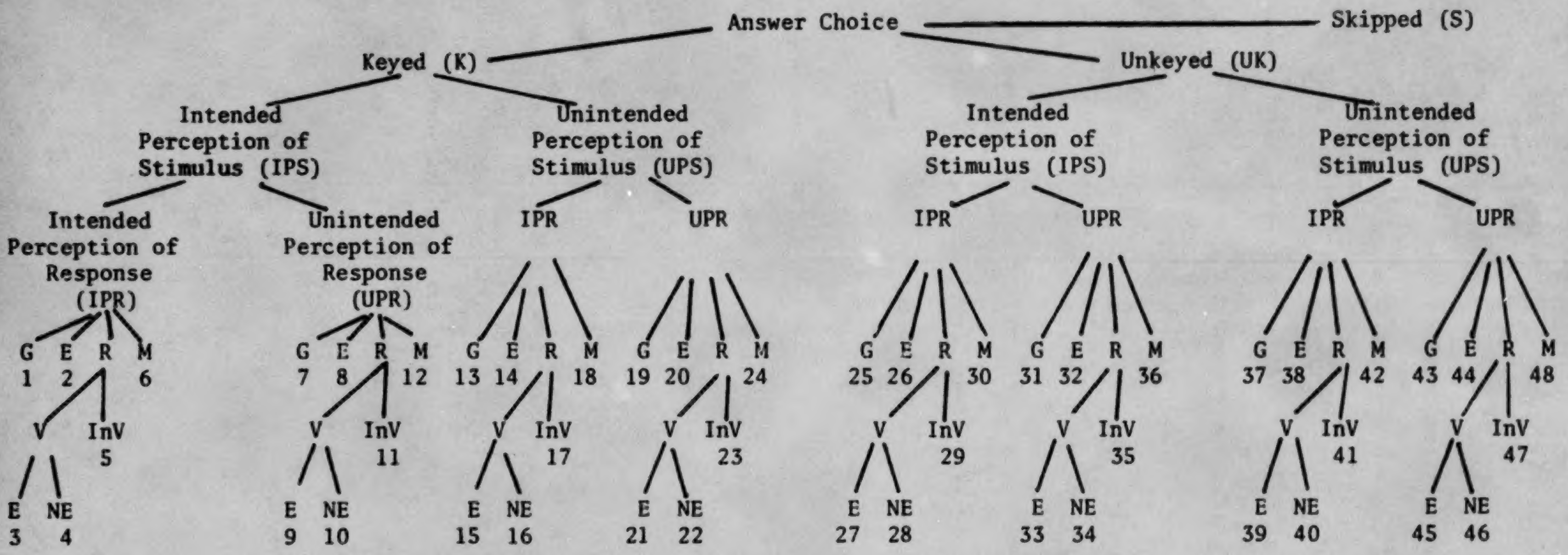
Social Studies. Test series' manuals also describe the content of the social studies subtests. The MAT Social Studies subtest is designed to test knowledge, comprehension, inquiry skills, and critical analysis of geography, history, economics, sociology, political science, anthropology, and psychology (Prescott, Balow, Hogan, & Farr, 1978). The SAT is said to measure achievement in geography, history, economics, political science, sociology and anthropology (Madden et al., 1973). The CTBS asserts that its items "measure the student's grasp of concepts, generalizations, and inquiry skills necessary for effective problem solving in social studies." These skills are tested in settings drawn from four content areas: the physical environment, the social environment, the political/economic environment, and history" (CTB/McGraw-Hill, 1976, p. 5). As with science items all social studies items are read aloud by the examiner; pupils select their answers from among three or four drawings.

Raters were familiarized with these statements of what each type of item ostensibly measured, and to help them keep this in mind when they were rating children's accounts of how they answered items, specific content and skill designations for each item were recorded on the rating forms they used. Also, prior to rating of the transcripts it was indicated on rating forms whether a child had marked keyed, unkeyed alternatives or had skipped the item in taking the test. This was done because this aspect of rating was simply clerical, calling for no critical judgment.

Other ratings were of course not so straightforward. Ratings pertaining to children's perceptions of item stimuli and response alternatives, the process by which they reached answers, the relevance of their logic to the attribute ostensibly measured, and whether or not the child employed extraneous knowledge, all called for considerable judgment on the part of the raters. The last two rating categories were pertinent only if the rater judged the child to have applied logic or reasoning in the process of selecting an answer.

The entire process of analytically rating an interview transcript regarding a single item is depicted in Figure 1, in the form of a decision tree. In terms of this decision tree there are 48 different possible outcomes for a complete

FIGURE 1, DECISION TREE FOR ANALYTICAL RATINGS



KEY:	
G	Guess
E	Elimination
R	Reasoning
M	Mistake
V	Valid
InV	Invalid
E	Child Used Extraneous Knowledge
NE	Child Did Not Use Extraneous Knowledge

rating of any one child-item interview. These are numbered for easy reference. Rating outcome 4, for example, would indicate that the child's reasoning followed the following sequence: keyed answer (K), intended perception of stimulus (IPS), intended perception of response alternative (IPR), reasoning employed in deriving an answer (R), reasoning valid with respect to the attribute ostensibly measured (V), and no extraneous information used in deducing an answer (NE).

Application of this rating scheme obviously raises questions of reliability since raters are asked to draw some very subtle judgments. Prior to ratings, a set of written instructions regarding use of the rating scheme was prepared and discussed with both raters in order to clarify ambiguities concerning rating categories. An initial tryout of the rating scheme, by two independent raters on the same child's interview transcript, showed that the ratings had only moderate inter-rater reliability -- less than 70% agreement. The rating scheme was again discussed with the raters and after rating instructions were refined the raters again were asked independently to rate the same interview transcript. This second tryout produced an inter-rater consistency of 78%. Inspection of items on which raters disagreed suggested that in some cases raters had simply made clerical errors or had misinterpreted particular response categories; for example, one rater had misunderstood the point that if a child showed an unintended perception of a single alternative answer in a way unrelated to the skill ostensibly measured, then the rating should be unintended perception of response alternatives (UPR) rather than intended (IPR). These points were brought to the attention of raters and they were asked to check over their second tryout ratings both for clerical errors and in light of a careful review of rating instructions.

After these self-corrections by raters, the inter-rater consistency on the second tryout increased to 90% -- a level of consistency which seemed reasonably acceptable.

Both raters then independently rated each transcript for each subtest for each child (excepting the social studies transcript for child B which was used in training the raters). After the ratings were corrected for clerical errors, inter-rater consistency for the reading comprehension, science and social studies subtests were 90%, 88% and 90% respectively.[*] For the purposes of analysis reported in the next section, items on which raters disagreed were resolved through discussion between the two raters and one of the investigators. Nevertheless, since these cases of disagreement reveal several important points about the inadequacies of the rating scheme, and also about the subtleties of drawing inferences about children's reasoning, we shall come back in section V, discussion of findings, to recount some of the apparent causes of disagreement.

4.2 Analytical Rating Results

Findings from the analytical ratings are presented in two parts. First described are aggregate results of analytical ratings of children's accounts of how they answered particular test items, and then the detailed results of the analytical ratings.

---

[*] It should be noted that the consistency data cited here and above are based on items answered (that is excluding skipped items) and refer to consistency over the total set of ratings for items (that is consistency in where item ratings came out at the bottom of the decision tree shown in Figure 1). Thus consistency for individual ratings generally were somewhat higher than 90%. More discussion on this point is provided in section V.

Aggregate Results of Analytical Ratings

Table 3 presents the results of analytical ratings of transcripts of interviews with children about how they answered items on the composite tests. Ratings pertain to the five categories under which raters were asked to make judgments concerning transcripts of interviews with children, as explained above. As depicted in the decision tree in Figure 1, the last two categories of rating were contingent on previous ones.

In only a single instance was a child rated to have perceived an item stimulus in a way that was unrelated to the skill ostensibly measured by the item. Unintended perceptions of response alternatives were more common -- in 5% of the cases for science items and 21% of the cases for social studies items. The findings of zero cases of unintended perceptions of response alternatives for reading items resulted at least in part from the point noted above -- namely, that it would be hard for children's reading aloud of reading item response alternatives to reveal unintended perceptions unrelated to the skill of reading comprehension.

Ratings of the process or manner in which children picked response alternatives indicated that in the vast majority of cases students evidenced some direct form of reasoning or logic in selecting answers (91% for reading comprehension, 88% for science and 89% for social studies). The second most common means of deriving answers appears to have been guessing (in 5% of the cases for reading comprehension, 8% for science and 10% for social studies).

TABLE 3: AGGREGATE RESULTS OF ANALYTICAL RATINGS OF ITEM-INTERVIEWS

	Reading Comprehension*	Science	Social Studies
Total No. of Item Interviews	358	319	232
Total No. of Interviews Concerning Items Not Inadvertently Skipped	327	299	224
Perception of Item Stimulus			
Intended (IPS)	NA	299 (100%)	224 (100%)
Unintended (UPS)		0	0
Perception of Item Response Alternatives			
Intended (IPR)	327 (100%)	285 (95%)	177 (79%)
Unintended (UPR)	0	14 (5%)	48 (21%)
Process			
Guess (G)	18 (5%)	23 (8%)	22 (10%)
Elimination (E)	4 (1%)	1 (0%)	1 (0%)
Mistake (M)	8 (2%)	11 (4%)	2 (1%)
Reasoning (R)	297 (91%)	264 (88%)	200 (89%)
Reasoning (For R items only)			
Valid (V)	237 (80%)	192 (73%)	136 (68%)
Invalid (InV)	60 (20%)	72 (27%)	64 (32%)
Information (for V items only)			
Extraneous Info. Used (E)	4 (2%)	4 (2%)	1 (1%)
No Extraneous Info. Used (NE)	233 (98%)	188 (98%)	135 (99%)

---

[*] Since children were not systematically asked to read reading comprehension passages, ratings concerning perception of item stimulus were not applicable to reading comprehension items.

In a small number of cases (2% of reading items, 4% for science items and 1% of social studies items) children had simply made a mistake in marking their answers. Instances were rated as mistakes only when children spontaneously changed their answers in the course of the interview, without prompting from the adult interviewer. In still fewer cases were children rated to have employed some form of elimination in selecting response alternatives (in 0-1% of the cases for the three subtests).

In cases where children were rated to have employed some direct form of reasoning in answering test questions, a subsequent rating pertained to validity of children's reasoning in light of their perception of item stimulus and response alternatives and in view of the attribute ostensibly measured by the item. The attribute ostensibly measured may have been validly demonstrated even though the child selected an unkeyed response alternative.

The InV rating category was intended to designate reasoning which was not a reflection of the attribute ostensibly measured, even though such reasoning might be interpreted as valid in terms of other attributes or points of view. In short a child's reasoning could be rated invalid regardless of whether the child picked the keyed or the unkeyed response alternative. In the aggregate, children's reasoned responses were rated invalid in 20% of the reading comprehension items, 27% of the science items, and 32% of the social studies items.

The last rating category, extraneous knowledge used, pertained only those cases in which children were rated to have employed reasoning validity reflecting the attribute ostensibly measured. In such cases raters were asked to indicate whether or not children in their reasoned accounts of how they selected response alternatives employed extraneous information not actually given in the item stimulus. Ratings indicated such use of extraneous information to be relatively rare -- in only 1 to 2% of the cases rated to show reasoning validly reflecting the attribute ostensibly measured.

Detailed Results of Analytical Ratings

Detailed results of ratings, in terms of final rating categories of the decision tree depicted in Figure 1, are given in Table 4. What do these ratings mean? In particular, how might we go about defining ambiguity in test items in terms of these ratings? Let us begin by trying to define this condition (and to estimate its prevalence) by noting its absence.

Cases of Clarity. To avoid a linguistic infelicity, we refer to cases of clarity in children's encounters with test items, rather than to "unambiguities." As with ambiguity there are two distinct forms of clarity in children's encounters with test items. First children may select a "correct," or in our terms keyed, answer to a question while actually applying the skill or attribute which the test item ostensibly measures. Second they may choose an "incorrect" or unkeyed answer while failing to demonstrate, at least in some respect, the skill or attribute ostensibly measured.

In terms of our rating scheme, these two types of clarity in children's encounters with test items may be defined as follows: First keyed answer clarity occurs when the child chooses the keyed response alternative,

TABLE 4: NUMBER AND PERCENTAGE OF ITEM INTERVIEWS IN EACH ANALYTICAL RATING CATEGORY

Rating Category # (see Fig. 1 for explanation)[1]		Reading Comprehension		Science		Social Studies	
		No.	%	No.	%	No.	%
	1	7	2.0	11	3.4	10	4.3
	2	2	0.8	1	0.3	1	0.4
	3	3	0.8	3	0.9	1	0.4
	4	231	64.5	182	57.0	111	47.8
Keyed Response Selected	5	14	3.9	9	2.8	12	5.2
	6	0	0	1	0.3	0	0
	7	0	0	1	0.3	4	1.7
	10	0	0	5	1.6	20	8.6
	11	0	0	1	0.3	3	1.3
	12	0	0	0	0	1	0.4
	17	0	0	0	0	1	0.4
	25	11	3.1	9	2.8	7	3.0
	26	1	0.3	0	0	0	0
	27	1	0.3	0	0	0	0
	28	1	0.3	0	0	0	0
Unkeyed Response Selected	29	46	12.8	60	18.8	34	14.6
	30	8	2.2	9	2.8	0	0
	31	0	0	2	0.6	1	0.4
	33	0	0	1	0.3	0	0
	34	1	0.3	1	0.3	4	1.7
	35	0	0	2	0.6	14	6.0
	36	0	0	1	0.3	1	0.4
Skipped		31	8.6	20	6.3	7	3.0
	Total	358	99.9	319	99.6	232	99.6

---

[1] No item interviews were rated as belonging in categories 8,9,13,14,15,16, 18,19,20,21,22,23,24,32,37,38,39,40,41,42,43,44,45,46,47,48.

shows the intended perception of item stimulus and response alternatives, evidences reasoning validity related to the attribute ostensibly measured, and does not apply extraneous information in the reasoning process. In terms of notation used in ratings, this could be summarized as K-IPS-IPR-R-V-NE, leading to final categorization number 4 as shown in Figure 1. Clarity with respect to unkeyed answers can be defined to refer to cases in which the child picks an unkeyed answer, while perceiving item stimulus and response alternatives in intended fashion, but without demonstrating valid application of the skill ostensibly measured (that is the process being guessing or invalid reasoning). In terms of our notation this corresponds to UK-IPS-IPR-and G or R-InV, chains of rating designated by numbers 25 and 29 in Figure 1.

The proportions of item accounts so rated are given in Table 5. Given that clarity or lack of ambiguity is defined by these outcomes of our rating scheme as applied to transcribed accounts of how children reasoned about test items, it can be estimated that no ambiguity was apparent in 90%, 85% and 68% of the cases in which children picked keyed answers on reading comprehension, science and social studies questions respectively. Similarly in cases in which children picked unkeyed response alternatives, no ambiguity was apparent in terms of our rating scheme in 83%, 81% and 67% of the reading comprehension, science and social studies item interviews respectively. Summing across both keyed and unkeyed answers, clarity as defined above was apparent in 88%, 84% and 68% of the three types of items, if we exclude items which children skipped (or 80%, 79% and 66% of the cases if we include instances in which children skipped items while actually taking the composite tests).

TABLE 5: CASES OF CLARITY IN ANALYTICAL RATING RESULTS

		Subtest		
		Reading Comprehension	Science	Social Studies
Total No. of Item Interviews		358	319	232
Keyed		258	214	164
Unkeyed		69	85	61
Skipped		31	20	7
Clarity in Keyed Answers				
Rating Chain	Rating Outcome in Fig. 1			
K-IPS-IPR-R-V-NE				
No.	1	231	182	111
% of Keyed answers		89.5	85.0	67.7
Clarity in Unkeyed Answers				
Rating Chain	Rating Outcome in Fig. 1			
UK-IPS-IPR-G	25			
-R-InV	29			
No.		57	69	41
% of Unkeyed Answers		82.6	81.2	67.2
Clarity - Total				
No.		288	251	152
% of Total (excluding skipped items)		88.1	84.0	67.6
% of Total (including skipped items)		80.4	78.7	65.5

Cases of Ambiguity. One might attempt to define ambiguity simply in terms of complements of the foregoing percentages, that is by subtracting the above percentages from 100%. However, all that is not clear in terms of children's perceptions of test items is not clearly unclear. In other words, the boundaries between clarity and ambiguity of children's perceptions of test items are themselves rather fuzzy. Therefore let us here attempt to define such ambiguity more directly.

As suggested in the literature reviewed in section II, ambiguity is of two sorts: first, children may choose the keyed answers without applying the skill or attribute ostensibly measured; and second, children may demonstrate the skill or attribute ostensibly measured, even though they select an unkeyed or "wrong" answer.

The first instance, selection of keyed answers without clear demonstration of the attribute of interest, may be defined in terms of children's choosing a keyed answer by process of guessing (G) or mistake (M), or via direct reasoning which is not a valid reflection of the attribute of interest (InV). Numbers of cases rated as reflecting this condition are shown in Table 6.

The second instance, that is selection of an unkeyed response alternative, despite demonstration of the attribute of interest, occurs when children select an unkeyed response alternative even though they evidence reasoning which constitutes a valid representation of the skill ostensibly measured. Cases so rated are also shown in Table 6.

TABLE 6: CASES OF AMBIGUITY IN ANALYTICAL RATING RESULTS

		Subtest		
		Reading Comprehension	Science	Social Studies
Total No. of Item Interviews		358	319	232
Keyed		258	214	164
Unkeyed		69	85	61
Skipped		31	20	7
Ambiguity in Keyed Answers				
Rating Chain	Rating Outcome in Figure 1			
K-IPS-IPR-G	1	7	11	10
-R-InV	5	14	9	12
-M	6	0	1	0
K-IPS-UPR-G	7	0	1	4
-R-InV	11	0	1	3
-M	12	0	0	1
Total No.		21	23	30
% of Total Keyed Answers		8.1%	10.7%	18.3%
Ambiguity in Unkeyed Answers				
Rating Chain	Rating Outcome in Figure 1			
UK-IPS-IPR-R-V-E	27	1	0	0
-NE	28	1	0	0
UK-IPS-UPR-R-V-E	33	0	1	0
-NE	34	1	1	4
Total No.		3	2	4
% of Total Unkeyed Answers		4.3%	2.3%	6.6%
Ambiguity Total		24	25	34
% of Total Item cases (excluding skipped items)		7.3%	8.4%	15.1%
% of Total Item cases (including skipped items)		6.7%	7.8%	14.7%

Given these definitions of ambiguity in terms of the analytical ratings of transcribed accounts of how children reasoned about test items, it can be estimated that ambiguity was apparent in 8%, 11% and 18% of the cases in which children picked keyed or "right" answers to reading comprehension, science and social studies items respectively; and in 4%, 2% and 7% of cases in which children choose unkeyed or "wrong" answers on these same subtests. If these two forms of ambiguity are pooled, we can estimate that overall ambiguity was apparent in 7%, 8% and 15% of the cases for reading comprehension, science and social studies items respectively.

The estimates of percentages of cases of clarity and of ambiguity do not sum to 100%. The reason is simply that, as suggested above, the boundaries between clarity and ambiguity are themselves unclear. Instances in which children were rated to have employed some form of elimination, for example, have not been treated as cases of clarity or of ambiguity. Making such a distinction for cases of elimination would, we think, require examination on a case-by-case basis.

How many items were ambiguous? The foregoing estimates of cases of clarity and instances of ambiguity are based on the total number of child item interviews. A natural next question -- one which begins to address more directly the original purpose of the pilot study -- is how many items were actually involved in the cases of ambiguity identified. In other words, for how many of the items included in the composite tests was ambiguity of interpretation apparent for one or more of the students interviewed?

Data bearing on this question are presented in Table 7. As

TABLE 7: ITEMS FOR WHICH ONE OR MORE INSTANCES OF AMBIGUITY WERE APPARENT IN TERMS OF ANALYTICAL RATING

	Subtest		
	Reading Comprehension	Science	Social Studies
No. of Items	37	29	29
Average No. Correct (Keyed Answer Marked)	25.8	19.4	20.6
Average No. Incorrect (Unkeyed Answer Marked)	6.9	7.7	7.6
Average No. Skipped (No Answer Marked)	3.1	1.8	0.9
Items in Which Ambiguity Apparent in One or More Cases in Which Child Selected Keyed Answer,			
- Guess (Rating outcomes 1, 7)			
No.	7	8	9
% of all items	19%	28%	31%
- Reasoning Invalid (Rating Outcomes 5,11)			
No.	11	10	8
% of all items	30%	34%	21%
- Mistake (Rating Outcomes 6,12)			
No.		1	1
% of all items	0%	3%	3%
Items in Which Ambiguity Apparent in One or More Cases in Which Child Selected Unkeyed Answer (Rating Outcomes 27,28,33,34)			
Total No.	3	1	2
% of all items	8%	3%	7%

these data suggest, instances of ambiguity were apparent with respect to a moment range of items. Keyed answer ambiguity (that instance in which a child picked a keyed answer for reasons unrelated to the skill ostensibly measured) was apparent for as many as 30%, 34% and 31% of the items on the three composite tests respectively. As indicated in Table 7 most of this form of ambiguity derived from children's reasoning being invalid with respect to the attributes ostensibly measured. Unkeyed answer ambiguity (in which a child marked an unkeyed answer, even though his or her reasoning evidenced a valid application of the skill ostensibly measured) was apparent on 8%, 3% and 7% of the items on the three composite tests respectively.

Instances of ambiguity were fairly widely distributed across children as well as across items. At least one case of keyed answer ambiguity was apparent for 54% of the children taking the reading comprehension test, 83% for the science test, and 100% for the social studies test. At least one instance of unkeyed answer ambiguity was apparent for 27%, 16% and 37% of the children on the three composite tests respectively.

The greater occurrence of ambiguity with respect to keyed answer choices seems largely due to two factors. First, children in the pilot study chose keyed answers (answered "correctly") 72%, 67% and 71% of the items on the three composite tests respectively, whereas they chose unkeyed answers for only 19%, 27% and 26% of the items respectively. Second, recall that our definition of keyed answer ambiguity (in which a child picks a keyed answer without demonstrating the skill or attribute ostensibly measured by the item) includes instances in which children simply guessed a keyed answer, or marked it merely by mistake. Since guessing appears to be a fairly frequent strategy by which children choose answers (see Tables 3 and 7) this accounts in part for the apparently greater frequency of ambiguity with respect to keyed answers than with respect to unkeyed answers. Before commenting further on these results, let us first recount in the next section the second approach employed in rating children's accounts of how they reasoned about test items.

4.3 Holistic Ratings

The analytical rating process described in the last few sections entailed several problems. First, our rating scheme required extracting a greal deal of information from relatively brief interviews with children; specifically regarding ratings of children's perception of both item stimuli and response alternatives; of the process they used in selecting an answer; if reasoning was used, whether it was a valid reflection of the skill ostensibly measured; and also whether or not a child employed extraneous information in selecting an answer. As the individuals engaged in the analytical ratings frequently pointed out, there was simply not enough information provided in some item interviews on which to base such ratings with much confidence.

The second problem was more subtle. In our initial work with transcripts of interviews with children, some of the things children said about particular items or about why they answered particular items seemed hard to understand. Why, for example, would a child who identified a carrot as the only root among several answer alternatives refuse to identify the carrot as the root of a green plant? As we worked with transcripts, and began to compare different children's perceptions to the same item, some of the points which previously had seemed inexplicable, became easier to understand or at least to begin to understand (we will return to discuss the carrot example in detail in section V).

For these reasons, we decided to try a second approach to rating the transcripts, an approach which we call a holistic rating. For this approach we cut apart copies of transcripts of conversations with each child and put together item sets for each item on all three composite tests. Each item set contained a description of the item stimulus including appropriate accompanying illustrations, the item question and response alternatives, the distribution of answers selected, plus portions of the interview transcripts in which each child talked about that item. Thus, each item set contained the item and accounts of how approximately ten children perceived and reasoned about that item (specifically ten children for the reading comprehension items, eleven for the science items, and nine children for the social studies items).

The question addressed with these item sets, using the holistic rating approach, was whether ambiguity was apparent in how one or more chilren perceived and reasoned about each item. Two raters were employed in making the holistic ratings. Both were advanced doctoral students in developmental psychology and were instructed as follows: First we described the purpose and design of the pilot study. Second we explained our definition of ambiguity, namely that it was of two types: the first, unkeyed answer ambiguity occurs when a child marks an unkeyed answer, but whose description of how he or she answered the question indicates application of the skill ostensibly measured by the item; the second, keyed answer ambiguity occurs when a child marks the keyed answer while not demonstrating the skill ostensibly measured by the item. It was explained that the second type, namely keyed answer ambiguity could be of three different sorts: 1) a child might simply guess a keyed answer, 2) a child could mark a keyed answer by mistake, or 3) a child might mark a keyed answer for other reasons not reflecting the skill or attribute ostensibly measured. The two individuals asked to do the holistic ratings were familiarized with publishers, statements of what reading comprehension, science and social studies tests ostensibly measured, and examples of each sort of ambiguity were discussed with them. Then, the holistic raters were asked to use a common form, to indicate independently for each item set whether one or more forms of ambiguity were relatively clearly apparent for one or more children. It should be noted that the holistic raters were not informed of the results of the analytical ratings. The raters were also encouraged to offer written comments regarding how children apparently had perceived and reasoned about each item.

After independent ratings of the item sets were performed, results were compared. Two problems with respect to the holistic rating instructions were apparent. The first pertained to whether or not children had selected keyed or unkeyed answers. Answer alternatives selected by each child had been marked on the child's item interview transcript. In a few instances, children decided during the interview on an answer different than the one they had marked during the regular test administration. In such cases, both original answer selected and the answer to which the child changed during the interview had been marked on the transcript. In eleven instances, holistic raters had judged children's accounts in terms of their changed answer rather than their original answers. The second misunderstanding in holistic ratings pertained to the instruction that raters indicate only instances in which children's accounts clearly indicated one of the forms of ambiguity. However, because of uncertainty holistic raters marked twelve instances of ambiguity with question marks. In these instances, raters were asked to reconsider, and decide whether ambiguity was relatively clearly apparent, or was mainly conjectured. After these two obvious types of misunderstanding had been corrected, the inter-rater agreement between the two sets of holistic ratings ranged from 76% to 100% as shown in Table 8.

4.4 Holistic Rating Results

Results of the holistic ratings are shown in Table 9. In light of a moderate level of disagreement between the two independent holistic ratings of item sets, we have taken a conservative approach in presenting these results, namely by summarizing results in terms of the cases in which the independent raters agreed that a particular form of ambiguity was apparent in how one or more children perceived and reasoned about an item.* Unkeyed answer ambiguity was apparent for at least one child on 22%, 21% and 14% of the reading comprehension, science and social studies items respectively. Keyed answer ambiguity, at least of two sorts, was more prevalent. At least one child was rated by both holistic raters to have merely guessed a keyed answer on 11%, 31% and 41% of the three item types respectively. Children were rated to have marked keyed answers by mistake for far fewer of the items -- 0%, 0% and 7% with respect to the three item types. At least one child was rated to have selected keyed answers for other reasons, without demonstrating the skill ostensibly measured for 30%, 24% and 38% of the reading comprehension, science and social studies items. Examples of explanations given by holistic

---

*Note that this approach to summarizing findings was not possible with respect to the analytical ratings because of the contingent nature of the analytical ratings.

TABLE 8: INTER-RATER AGREEMENT FOR HOLISTIC RATINGS

Type of Ambiguity Identified	Composite Subtest		
	Reading Comprehension	Science	Social Studies
Unkeyed Answer Ambiguity In which one or more children selected unkeyed or incorrect answer while demonstrating skill ostensibly measured)	76%	90%	100%
Keyed Answer Ambiguity (In which one or more children selected a keyed answer without demonstrating skill ostensibly measured via one of following)			
Guess	86%	90%	83%
Mistake	100%	97%	93%
Other Process	76%	86%	76%

TABLE 9: RESULTS OF HOLISTIC RATINGS*

	Composite Subtest					
Type of Ambiguity	Reading Comprehension		Science		Social Studies	
	No.	% of Total	No.	% of Total	No.	% of Total
Keyed Answer Ambiguity (In which one or more children selected a keyed answer without demonstrating skill ostensibly measured via one of following)						
Guess	4	11%	9	31%	12	41%
Mistake	0	0%	0	0%	2	7%
Other Process	11	30%	7	24%	11	38%
Unkeyed Answer Ambiguity (In which one or more children selected unkeyed or incorrect answer while demonstrating skill ostensibly measured)	8	22%	6	21%	4	14%

---

*Results given here indicate items on which particular forms of ambiguity were clearly apparent to both of the holistic raters.

raters with respect to this latter category, that is other bases on which children selected keyed answers, were as follows:

Reading Comprehension Items	- "uses personal experience"
	- "makes up own story"
	- "selected plausible answer"
	- misperception "carryover from previous question"
	- "without reference to story [child] personalizes"
Science	- "elimination"
	- "thinks fish breathe water"
	- "used pre-existing knowledge"
	- "makes up story to explain choice"
	- "reason false"
Social Studies	- "reasons by counter example"
	- "uses other knowledge, not time line [given]"
	- "thinks south pole gets the least heat and north pole the most"
	- "personal knowledge about Hawaii made him choose answer"

These are simply a few examples of explanations offered by holistic raters of how children selected keyed answers without demonstrating the skill ostensibly measured by the item. Many of these explanations are impossible to understand out of context, that is without first describing items, so let us postpone further discussion of them until section V, when we do discuss specific items.

4.5 Overall Findings

Having presented the results of the two different approaches to analyzing transcripts of conversations with children regarding how they perceived and reasoned about test items, we may now compare the broad patterns of results from these two approaches. Second, we examine the consistency of the two rating approaches in terms of items identified as shown to lead to ambiguity of each type for one or more children. Third, we consider how inferences may be drawn from the pilot sample of items to the broader population of items from which our sample was drawn.

Patterns of Results. Results of holistic ratings presented in Table 9 may be compared with results of analytical ratings presented in Table 7. First note some of the broad similarities in the outcomes of these two different and independent approaches to rating children's accounts of how they perceived and reasoned about items. The ranges of proportions of items of each type identified as evidencing various sorts of ambiguity for one or more children were approximately equal -- ranging from 0% to 34% in the analytical ratings and 0% to 41% in the holistic ratings. Frequencies with respect to types of keyed answer ambiguity are also roughly comparable in the two approaches. Selection of a keyed answer simply by mistake was rated to have been relatively rare for each of the three types of items in both rating approaches. Guessing of a keyed answer was considerably more common, occurring in 19% to 31% of items according to the analytical ratings and 11% to 41% of items according to holistic ratings. Also, under both rating approaches, keyed answer ambiguity for reasons other than guessing or mistakes was relatively common across all three subtests; for 21% to 34% according to the analytical ratings and 24% to 38% according to the holistic ratings. Second, note that there is one major difference in the pattern of results across the two rating approaches. Specifically, unkeyed answer ambiguity (in which a child picks an unkeyed answer, but whose description of how he or she answered the item, indicates actual application of the skill ostensibly measured) was more prevalent in terms of the holistic ratings (rated as apparent for at least one child in 22%, 21% and 14% of the reading, science and social studies items respectively) than in the analytical ratings (occurring in only 8%, 3% and 7% of the same types of items respectively). The explanation for this discrepancy appears to derive from the fact that different phenomena were included under the "unkeyed answer ambiguity" rubric under the two rating approaches. Recall that under the first rating approach unkeyed answer ambiguity was defined strictly in terms of analytical rating outcomes 27,28, 33 and 34. As depicted in Figure 1, these outcomes corresponded to instances in which children exhibited the intended perception of the item stimulus, but who chose an unkeyed response alternative even though their reasoning was rated to constitute a valid reflection of the skill ostensibly measured by the item. The holistic raters did identify this type of phenomenon as unkeyed answer ambiguity, but they also included two other types of phenomena under this rubric: 1) cases in which children made mistakes in marking answers, and 2) instances in which children had misperceived item stimuli, but still seemed to demonstrate the skill ostensibly measured by the item. These interpretations seem quite plausible, especially since the same interpretation was drawn by independent holistic raters. Thus, our judgment is that the estimates of prevalence of unkeyed answer ambiguity shown in Table 9 may represent a more accurate estimation of the prevalence of unkeyed answer ambiguity than those in Table 7, which are implicitly based on a much narrower definition of unkeyed answer ambiguity.

Consistency of Results. Although broad patterns of results from the analytical and holistic rating approaches are somewhat similar, there was imcomplete consistency in the exact items identified via the two approaches as having evidenced ambiguity for one or more children. Table 10 shows the degree of consistency between the two approaches in terms of items identified as showing ambiguity of four different types for one or more of the pilot study sample of children. Two points may be noted with respect to these data. First, levels of consistency between results of different rating approaches are roughly comparable to the inter-rater reliability or consistency for two raters independently using each rating system (see pp. 30-31 and p. 48 , Table 8.) Second, the consistency between the two approaches was somewhat greater with respect to cases in which children picked keyed answers merely as a result of a guess or a mistake (in the latter occurrence, recall was relatively rare) than with respect to cases of unkeyed answer ambiguity (in which children picked unkeyed answers, but were rated to have demonstrated the skill ostensibly measured or to have reasoned validly in light of their perceptions of item stimuli and response alternatives) or keyed answer ambiguity deriving from other than guessing or making a mistake in marking answers.

Statistical Significance of Results. Up to now, results of our analysis of ratings of transcripts of conversations with children regarding how they perceived and reasoned about test items have been presented merely in descriptive terms; that is without calculating any tests of statistical significance. Nevertheless, a relevant question is how we can draw inferences from the particular random sample of items included in the pilot study back to the original pool of items from which the pilot study items were drawn.

TABLE 10: CONSISTENCY IN ITEMS IDENTIFIED AS AMBIGUOUS BY TWO RATING APPROACHES.

	Composite Subtest		
Categories of Ambiguity	Reading Comprehension	Science	Social Studies
Unkeyed Answer Ambiguity	81%	76%	93%
(Analytical Rating Outcomes) 27, 28, 33, 34			
Keyed Answer Ambiguity			
Guess (Analytical Rating Outcomes 1, 7)	92%	83%	90%
Mistake (Analytical Rating Outcomes 6, 12)	100%	97%	90%
Holistic rating "other" -- Analytical Rating "Reasoning Invalid" (Outcomes 5, 11)	84%	83%	62%

The standard method for calculating confidence limits for a proportion is as follows (see Snedecor & Cochran, 1972): If m members out of a sample size n are found to possess some attribute (in our case ambiguity), the sample estimate of the proportion in the population possessing this attribute is $p = m/n$. In large samples (our sample was not very large, but this is a point to which we will return in a moment), the estimate p is approximately normally distributed about the population proportion P with standard deviation $\sqrt{PQ/n}$. Since the population proportion P is unknown, the sample estimate $\sqrt{pq/n}$ may be substituted for the true but unknown standard deviation $\sqrt{PQ/n}$. Hence the probability is approximately .80 that P lies between.

$$p \pm 1.28 \sqrt{pq/n}$$

Since our sample sizes were all approximately 30, we can thus estimate with .80 probability ($z = 1.28$), that when $p = .10$, the true proportion P is greater than .03 but less than .17. In other words, we can be reasonably confident that where our estimates of ambiguity are around .10 that the corresponding proportion of ambiguous items in the population of items from which our sample was drawn would be in the range of .03 - .17.

Several limitations in this estimate of confidence intervals should be noted; concerning 1) the confidence level of .80; 2) our sample size; 3) redundancy of estimation on a single sample; and 4) the appropriateness of inferences, not just across items, but across children. First, regarding confidence levels, we should note that we initially choose a confidence level of .80, rather than a more typical confidence level of .90 or .95, both because our inquiry was only a pilot study and because we wanted to keep our sample size of items to a reasonable number and not subject the pilot sample children to too great a number of items (also, of course, we had no indication beforehand of the proportions p of ambiguity which we would find, so in planning a sample we used the conservative estimate of 0.5). However, it should be noted that if we were to increase the confidence level in our calculations of confidence intervals to .90 or .95 (with corresponding Z values of 1.64 and 1.96) that our confidence intervals would increase 28% and 53% respectively.

Second, the method of calculating confidence intervals for proportions cited above is, strictly speaking, for large samples only, where p is not close to 0 or 1.00. Since our sample was relatively small it would be more appropriate to employ a confidence interval for proportions based on small samples. The effect of such an approach would be to move the confidence interval upward. The confidence interval for $p = .10$, $n = 30$, confidence level .95, for example is .02-.26, whereas the large sample approximation to this is $.10 \pm .11$ (see Freund, 1972). Thus the confidence interval extimate above is somewhat biased in a downward direction (for $p < .50$).

Third, it should be noted that since we only drew three random samples of items (one each for reading comprehension, social studies and science); it is not, strictly speaking, appropriate to calculate more than three tests of statistical significance on our findings. For all of the foregoing reasons we will not belabor the issue of formal statistical inference and calculate estimates of statistical significance separately for each of the proportions given in Tables 7 and 9.

Finally, we should note that given the design of the pilot study, statistical inference applies to generalizations across items, but does not apply formally to generalization from the pilot study sample of children to any broader population of children. Thus statistical inference, strictly speaking, refers to the levels of ambiguity which likely would have been apparent had we administered the total population of items from which our sample was drawn to the same sample of children who participated in the original pilot study.

Generalization of our findings beyond this point depends not on tenets of formal statistical inference, but instead on plausibility and logical analysis. Before discussing such issues in section VI we think it will be useful to discuss what seem to have been the main sources of ambiguity in test items. Hence in closing section IV, let us simply state our narrowly defined findings, based on four independent ratings (using two separate rating schemes) of transcribed accounts of how children perceived and reasoned about a random sample of items. Specifically we estimate with 80% confidence that had our sample of children been given the total reading comprehension, science and social studies test forms from which the sample items were drawn, ambiguity of the following forms would have been apparent for one or more children for the following percentages of items*:

---

* A final note of clarification should be offered with regard to exactly how these estimates were derived. Since the confidence intervals for estimates of ambiguity of various sorts would overlap for the three types of items, it is reasonable to pool estimates across item types. We therefore averaged comparable estimates across the results of the two rating approaches given in Tables 7 and 9, and across the three item types, then calculated confidence intervals based on a .80 confidence level in the manner described above. Note that these estimates are conservative in that they are based on the large sample approximate method for calculating confidence intervals around proportions.

Keyed Answer Ambiguity	
- Choosing keyed answer by guessing	17 - 37%
- Choosing keyed answer by mistake	0 - 5%
- Choosing keyed answer for other reasons, but without demonstrating skill ostensibly measured	19 - 41%
Unkeyed Answer Ambiguity	6 - 20%

V. DISCUSSION: LIMITATIONS OF STUDY AND SOURCES OF AMBIGUITY

The calculations reported in the last section may lend an unwarranted sense of precision and finality to the findings of the pilot study Therefore before we go any further in discussing our findings, let us step back a moment to discuss some of the limitations of this approach to studying the qualities of test items and children's interpretations of them. Specifically in this section we elaborate on problems of 1) the interview process, and 2) the rating schemes and their application. After elaborating on these points, we discuss particular sources of ambiguity with respect to test items, and discuss the problem of defining ambiguity with respect to test items.

5.1 Interviews

Talking with children about how they perceive and answer test questions certainly allows us more insight into the processes by which they make sense of the questions than does mere inspection of the answers they mark. Nevertheless, such interviews are still a very imperfect means of gaining access to children's perceptions and reasoning. In this respect, we must re-emphasize that our interviews with each child concerning a particular item were extremely brief, lasting on average only a minute or two, and our interviewers were not specially trained in conducting clinical interviews with children. In contrast, specially trained research psychologists sometimes devote hours, or even days, to probing the nature of a child's reasoning with respect to just a single task.* Also, of course, simply asking children how they make sense of test items may not be an efficient means of ascertaining certain aspects of their perceptions and reasoning. Not only are young children generally less capable than adults in articulating their perceptions and reasoning, but some psychologists have offered evidence to suggest that even with adults "there may be little or no direct introspective access to higher order cognitive processes" (Nisbett & Wilson, 1977). Hence, our brief conversations with children about how they perceived and reasoned about test questions can be considered no more than crude inquiries into how they actually perceived and reasoned about test items. Nevertheless, this point does not, we think, detract from the significance of our findings. With relatively crude means of inquiry via brief interviews we have identified significant levels of ambiguity in children's perception and reasoning about a random sample of items. Had our means of inquiry been more refined, it seems reasonable to expect that ambiguities revealed would have been no less, and may even have been more.

5.2 Ratings

As students of nonverbal communication suggest, considerable information concerning social perception is lost when we move from face-to-face conversations to audio-recordings of the same conversations. Still more information may be lost when audio-recordings are transcribed onto the printed page -- as has often been observed with respect to the Watergate tapes.

---

* A sobering example of how difficult it is to discover how children perceive and reason about even as simple a task as building a bridge out of blocks is offered by Luchins & Luchins, 1977. Their account shows how several permutations on the block-building task are required simply to begin to understand how a child perceives and reasons about such a concrete task.

Nevertheless, probably a more severe diminution of information and insight into childrens encounters with test items treated in the pilot study occurred in going from transcripts of our conversations to ratings of them. Though conversations concerning each item were brief, transcripts of conversations conducted in the pilot study run to more than 300 pages altogether. To provide more insight into these accounts of children's perceptions, we will in a moment quote directly from those transcripts. First, however, let us recount some of the problems and weaknesses of the two rating systems employed.

Analytical Ratings. The analytical rating scheme was developed in an attempt to analyze children's accounts of how they answered test questions in a systematic manner. This was necessary to provide a basis for generalizing findings from the pilot study. As recounted in section 4.1, several trials of the rating system and refinements of instructions to raters were necessary before raters could independently use it and derive reasonably consistent ratings of the same transcripts. For the sake of analyses reported in section IV, inconsistent analytical ratings were resolved through discussion between the two raters and one of the investigators.

Nevertheless, cases in which independent analytical ratings were inconsistent are worth describing for what they reveal both about our rating process and about how hard it is to tell what it is that a particular item represents for a particular child. In analytical ratings there were 93 cases altogether, across all three composite subtests, in which independent ratings disagreed in terms of the final outcome rating category shown in Figure 1. In all but two of these cases the inconsistency in final rating category derived from differing judgments concerning a single level in the rating scheme. By far the most common source of rating inconsistency (in 42 of the 93 inconsistent cases) were judgments concerning validity of reasoning, that is, whether or not a child's account of how he or she answered a question did or did not reflect a valid demonstration of the attribute ostensibly measured by the item (ratings of V or InV).

Item 21 from our composite science test, for example, was the source of analytical rating inconsistency in four cases. This item presented children with a picture of the sun and earth with top and bottom of the earth marked N and S respectively. Children were verbally given the statement "When it's nighttime there's a full moon shining" and then were supposed to indicate "yes," "no" or "can't tell" from the picture what the answer is. The answer indicated to be correct in the publisher's scoring key is "can't tell." Child A marked "can't tell" and explained his reasoning as follows:

A: Can't tell.

L: Why not?

A: Cause sometimes, sometimes it's half the moon and sometimes it's full moon and quarter moon . . . I don't know.

According to the publisher this sort of science item measures "knowledge, comprehension, inquiry skills, and critical analysis in physical science, earth and space science and life science" (Prescott et al., 1978). One rater indicated that child A's response indicated a valid demonstration of these attributes, while the other rater initially indicated that it did not. The apparent source of this rating inconsistency was that the second rater judged that child's response did not demonstrate the science inquiry skills apparently implicit in this item. However, upon discussion, raters agreed that the child's response did reflect a valid knowledge of earth and space science even if the child's answer was not clearly based upon the illustration accompanying this item. Thus, the inconsistency in this rating stemmed as much from ambiguity concerning what it was that the item ostensibly measured as it did from ambiguity about how the child reasoned about the item. Because of such uncertainities, ratings made in the pilot study do themselves carry some degree of ambiguity.

Also, we should explain one other aspect of the analytical rating scheme which clearly affected our findings. Recall that in our rating system, raters were asked to indicate whether children perceived item stimuli and response alternatives in intended or unintended fashion. The unintended rating categories (UPS and UPR) were used to indicate that a child perceived either stimulus or response alternatives in an unintended fashion which was not directly related to the skill ostensibly measured by the item.

Recall also that we did not systematically ask all children to read reading comprehension passages prior to talking with them about how they answered reading comprehension test items. We did not do this for two reasons, one practical and one substantive. From the practical point of view, having children read passages out loud would have greatly increased the time necessary for interviewing. From the substantive point of view we had reasoned that there would be no way in which having children read passages out loud could reveal misperceptions which were unrelated to the skills of reading comprehension. This assumes, of course, that specific knowledge of vocabulary used in reading passages is a component skill of reading comprehension. This assumption led us to forego rating of children's perception of reading comprehension item stimuli, and also led directly to the relatively low frequency in ratings of unintended perception of response alternatives (UPR) with respect to reading comprehension. The assumption still seems a reasonable one to us. Nevertheless, it should be noted that we did find instances in which misunderstanding of a single word in either reading passages or response alternatives led to children's selecting unkeyed answers on reading comprehension items even though they seemed to comprehend most of the relevant passage. One boy, for example, did not grasp the fact that the word "Mouser" in a story about a cat was the cat's name, and failure to understand this single word prevented him from answering two of four items pertaining to the passage about Mouser -- even though he seemed to comprehend almost all other aspects of the reading passage.

Holistic Ratings. The holistic rating system was developed in order to try to overcome some of the problems in the analytical rating approach. The second strategy for analyzing transcripts of conversations with children attempted, for example, to reduce the number of judgments which raters were called on to make. In the analytical ratings, raters were asked to make as many as 4750 discrete judgments (95 items x 10 kids x up to 5 judgments). In the holistic ratings, this judgment load was reduced nearly 10-fold. Since the holistic raters were not asked to make judgments about how each child perceived and reasoned about each item, they were called on to make as few as 400 separate judgments (any of four types of ambiguity apparent for one or more children for 95 items). Nevertheless, despite the lower number of ratings called for in the holistic ratings there still were nearly 40 inconsistencies in judgments made by holistic raters. Although we attempted to minimize the effects of such inconsistencies in summarizing results of holistic ratings in Table 9, by counting cases as ambiguous only when both raters agreed in that designation, consideration of inconsistencies in holistic ratings helps to clarify why it is hard to judge precisely how children interpret test items. Question 12 on the composite science test was based on the following statement read aloud to children:

> Joan read that the surface of the Earth is about three fourths water. Three fourths of the Earth's surface is water. So she decided to draw a circle graph to show this. She used wavy lines for the water. What graph did she make?

Response alternatives for this question were three circles: the first was three-quarters wavy lines and one-quarter solid, the second was one-quarter wavy lines and three-quarters solid, and the third half-wavy lines and half solid. Child F selected the second answer alternative while taking the test under standard conditions. The interview with child F about this question went as follows. After repeating the item stimulus the interviewer asked, "Which graph shows what Joan would have made?"

F: I don't . . . second one.

L: Why is that?

F: Because she, it's like she said, I'm three-quarters of a half . . . .

L: Three-quarters of the . . . Earth . . .

F: And three-quarters is about that much. Three-quarters wouldn't be that much. Said most of the Earth.

L: Which one would say most of the Earth is covered?

F: Didn't it say most of the Earth?

L: It says about three-fourths of the Earth's surface is covered with water.

F: Say.

L: About three-fourths of the Earth is covered with water.

F: I don't get it. I did the opposite.

L: OK. You want to change your answer?

F: Yeah.

So child F went ahead and changed his answer to the first alternative, which was the keyed one.

One holistic rater judged this to be an instance of unkeyed ambiguity, on the grounds that the child had made a mistake and "corrects self." The other holistic rater did not indicate that this was a clear instance of ambiguity, but did comment that the child "may have the skill, but probing may have evoked change to keyed answer." One could plausibly make a case for either interpretation, but given our approach to summarizing the holistic ratings this was not counted as a case of ambiguity.

5.3 Sources of Ambiguity.

The point of the above examples is simply that there clearly was a certain degree of ambiguity in two important stages of the pilot study, namely the interview process, the two processes of rating interview transcripts. Therefore in this section let us more directly explain and illustrate some of the common sources of ambiguity in children's interactions with test items. Specifically we discuss examples from our interview transcripts as illustrations of the following sources of ambiguity in children's interactions with test items:

- rules of test-taking
- unusual and perceptive interpretation of questions
- personal experience and egocentric interpretations
- concrete sequential reasoning
- visual interpretation
- item format

It is impossible in most instances to say with certainty that ambiguity derived from any single one of these factors, but to illustrate what seems to have happened let us cite some specific examples.

Rules of test-taking. For each of the subtests, children were instructed to mark only one answer for each question. Since all children had taken standardized multiple-choice tests on previous occasions, according to their teacher, this was a rule of test-taking which might be assumed to be quite familiar to them. Yet, we found eight cases in which children seemed to forget this rule of test-taking and to mark more than one answer alternative.

Item number 23 on the reading comprehension composite test (MAT-2-1) was based on the following reading passage:

> Mother said, "Here is the cake I made. Would you like some?"
>
> Sue and Nick said, "Yes, thank you. We would."
>
> But their dog Red took Sue's cake and ran away with it. Mother got Sue some new cake. Then they all had cake to eat. Red too.

Question 23, based on this passage asked "Who made the cake?" Answer alternatives were:

- Sue
- Nick
- Mother
- Red

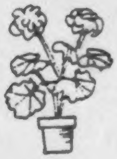
Nine out of ten children selected only the keyed alternative "Mother." Child J however marked both "Mother" and "Sue" and drew a line between the two. During our interview, child J explained his answers as follows:

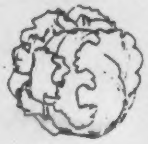
> I marked Mother and then I didn't think that made any sense because it could be anybody's mother, so I just wrote Sue's and then I drew a line to Mother, Sue's Mother."

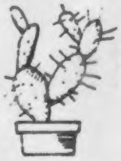
Clearly J comprehended the reading passage. His inferential comprehension seemed, if anything, more sophisticated than that reflected in the answer alternatives available. His solution to his dilemma was to mark two answers, transgressing the rule of marking only one, and hence to a response which would have been counted as incorrect under standard scoring procedures.

Unusual and Perceptive Interpretation. The above example was not, by any means, the only case of an unusually perceptive interpretation of a test item. Item 8 on our composite science test (SAT-20-13) consisted of the question "Which plant needs the least amount of water?" given orally with respect to the following three illustrations:

13


○


○


○

Children most commonly described these three illustrations as a "flower" or "some kind of plant"; "cabbage," "lettuce," or a "leafy vegetable," and "cactus," or "a prickle." Nine out of eleven children selected the keyed answer alternative, the cactus, as the plant needing the least amount of water.

Child A, however, selected the cabbage as the correct answer and explained his reasoning as follows:

> L: Which one needs the least amount of water?
>
> A: That does.
>
> L: Why is that?
>
> A: 'Cause it's a cabbage. Doesn't need as much water. Only when you clean it.

This child apparently reasoned that the cabbage needed less water than the two potted plants because it was not depicted as a potted cabbage. From this point of view, the logic of the child's reasoning is hard to deny. A head of cabbage once harvested can be kept well and long without water -- water is needed only when you clean it.

Personal Experience and Egocentrism. Another factor which clearly influenced children's perception and reasoning about test items was their own personal experience and egocentrism -- that is, their tendency to view things from their own personal point of view.

Children's reactions to item number 1 from our reading comprehension composite subtest (SAT-5-3) illustrate this tendency. This item was of the sentence completion type. The item stimulus was:

- Eve likes to watch TV. She thinks it is _____.

Answer alternatives were:

- fun
- dull
- hard
- petty

Nine out of ten children selected the intended correct answer "fun." Interestingly, however, all ten children read the last alternative not as "petty" but as "pretty." This illustrates a type of ambiguity which was not covered under our definition of ambiguity, but this is a point to which we will return in conclusion. For the moment let us simply consider two children's accounts of how they answered this question. The interview with child J went as follows:

L: Why don't you just start at number one, OK?

J: Eve likes to watch TV. She thinks it is fun.

L: What are the other words?

J: Dull, hard, pretty.

L: Hm-hum. And why did you pick the first one?

J: Because I think it's fun.

L: Oh.

J: I like to watch TV.

Clearly child J interpreted this question in terms of her own personal experience, not necessarily in terms of an inference that because Eve likes to watch TV she must think it fun.

Child H also seemed to answer this question in terms of his own personal experience, but it was a different personal experience.

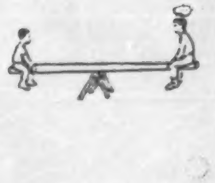
> W: OK, you want to just read this? You don't read the sample, just start on number one, OK.
>
> H: We are, we saw . . .
>
> W: No, that's the sample. Why don't we just start here?
>
> H: No, I want to start here.
>
> W: But we already talked about that in class, though. That's why I thought it wouldn't be necessary.
>
> H: Oh. Eve likes to watch TV. She thinks it is . . . fun, dull, hard or pretty. Dull.
>
> W: OK. Why did you pick that one?
>
> H: Because your eyes could get, us, your eyes could get, um, um, your eyes could get, you know, your eyes could get, you could get blind.
>
> W: Uh. . . . if you like to watch TV, you mean?
>
> H: Yeah, if you watch TV a lot.

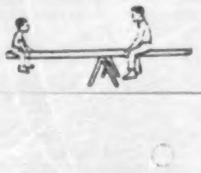
From his experience, this child reasoned that someone who likes TV, who watches it a lot will think it dull in the sense that watching TV a lot can hurt one's eyes. There are, of course, other grounds for thinking TV dull, but child H's interpretation of this question apparently stems from his own personal experience. The child seemed to understand the reading passage as well as other children, but his personal frame of reference in drawing inference from it was quite different, for example, then that of the girl who answered "fun" because she thinks TV is fun.

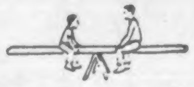
Another interesting example of how personal experience can affect children's perceptions of test items comes from child B's answer to item 26 on our composite science test (CTBS-30-23). The science item consisted of the following instructions and response alternatives:

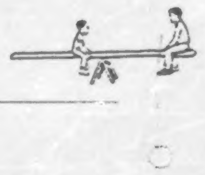
> There's a heavy boy and a lighter boy. How should they sit to balance on the seesaw?

23









Five of eleven children chose the keyed answer alternative, the second picture. Five selected the first and one gave no answer to this item. Child B explained his choice of the first picture as follows:

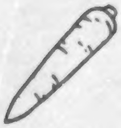
> B: Well, usually when my little brother and I get on the seesaw we sit at the ends because if I sit closer then we might -- he'll say, "get back, get back." He's afraid he'll fall.
>
> L: And is he bigger or smaller than you are?
>
> B: Smaller.

Concrete Sequential Reasoning. One could interpret the two examples concerning TV watching not only as showing that children interpret test items in light of their own experience, but also that their reasoning may be concrete and sequential. A child may be able to read the sentence "Eve likes to watch TV," but the next incomplete sentence is viewed separately, not necessarily in terms of Eve, but in terms of a separate consideration of what it is reasonable to think about TV. From this perspective one could, of course, argue that child H's answer in the example cited above exhibited concrete reading comprehension. The point seems debatable with respect to the case above, but there are other examples which even more clearly demonstrate children's tendency to think concretely and sequentially about the problems posed in test questions. Item number 14 from our composite science subtest (MAT-19-21) provides a particularly interesting example. The item stimulus, read aloud to children, was:

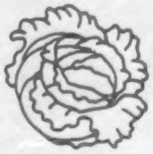
- Which one is the root of a green plant?

Answer alternatives were:

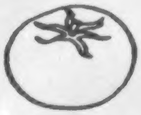
21



○



○



○

Children identified these pictures as "carrot" or "carrot without the top"; "cabbage," "lettuce," or "head of lettuce"; and "tomato," or "I don't know what that is." Despite the apparent ease with which most children identified these pictures, only three of eleven selected the keyed answer which was the picture of the carrot. Six had marked the cabbage or lettuce as the correct answer, and two had not marked any answer to this question. The interview with child D, who had not marked any answer went as follows:

L: Number fourteen. What are those pictures?

D: Carrot without the top, lettuce and I don't know what that is.

L: The last one?

D: Yes

L: O.K. Which of those is the root of a green plant? . . . What's a root?

D: It's the bottom of a plant.

L: Are any of those the bottom of a plant?

D: Ya.

L: Which one?

D: The carrot.

L: The carrot is. Is that the root of a green plant?

D: No! It's the root of an orange one.

Most of the children who had marked the cabbage or lettuce as the correct answer to this question, seemed to experience the same trouble as child D. They knew what a root is, knew that a carrot is orange, a cabbage (or lettuce) green, and a tomato red. But since there seemed to be no really correct answer to this question, they picked the cabbage picture as the best of several bad alternatives, since the question did clearly refer to a green plant. At one level, these children appeared to engage in a sort of concrete sequential reasoning. Since none of the answer alternatives seemed to make sense, they focused on a single part of the question, the reference to a green plant, and simply marked what clearly was a green plant.

At another level, an epistemological one, it could be argued that child D's response, namely a refusal to mark any of the response alternatives was the most thoughtful. After all, the root is a part of a plant -- the "bottom of a plant" as child D put it -- and if a plant has an orange root it is therefore not a green plant. Even though children may have thought in such terms it seems unlikely that they could verbalize them very well. But whatever the case it seems clear that some of the ambiguity in this item derived from the fact that children may think in concrete, sequential ways about answers to such problems.

Another reading passage and a child's interpretation of a question about it also suggests children's attention to concrete details, but also reveals real ambiguity with respect to the inferential comprehension expected of children. The reading passage, on which composite reading test questions 12-16 (CAT-10-15-19) were based, was as follows:

> Jane Addams cared about people all her life. When she lived in Chicago, she saw many children who did not have a good place to play. They played in the streets. Jane Addams thought the children were not safe. She wanted to do something for them.
>
> Jane Addams was given a large lot. The lot had many old, unused houses on it. She had the houses torn down so that the lot would be empty.

> Jane Addams worked hard to clean up the lot. Then she bought swings and seesaws. This became the first playground in Chicago. The children had a safe place to play!

Five comprehension questions were based on this passage. One of them, and the answer alternatives posed for it, was: What did Jane Addams do with the swings and seesaws?

- cleaned them
- played with them
- had them put on the lot
- used them in the streets

Six out of ten children marked the third answer alternative, which was the intended correct answer. Two marked "played with them" -- apparently reflecting the natural view of children about what one ought to do with swings and seesaws. One child, asked why he chose this answer alternative, explained "Because she could have lots more fun instead of sitting around in the house." One girl marked the first answer alternative. The conversation regarding her answer to the question went as follows:

> L: What about number fifteen?
>
> B: What did Jane Addams do with the swings and seesaws? Played with them, had them put in the lot, used them on the streets. I just said cleaned them.
>
> L: Why did you choose cleaned them?
>
> B: 'Cause she doesn't, she doesn't want the kids to go on dirty things.

According to the publisher's scoring key this is, of course, an incorrect answer. The reading passage does not say that Jane Addams cleaned the swings and seesaws -- but neither does it say that she sense, they focused on a single part of the question, the reference to a green plant, and simply marked what clearly was a green plant.

At another level, an epistemological one, it could be argued that child D's response, namely a refusal to mark any of the response alternatives was the most thoughtful. After all, the root is a part of a plant -- the "bottom of a plant" as child D put it -- and if a plant has an orange root it is therefore not a green plant. Even though children may have thought in such terms it seems unlikely that they could verbalize them very well. But whatever the case it seems clear that some of the ambiguity in this item derived from the fact that children may think in concrete, sequential ways about answers to such problems.

Another reading passage and a child's interpretation of a question about it also suggests children's attention to concrete details, but also reveals real ambiguity with respect to the inferential comprehension expected of children. The reading passage, on which composite reading test questions 12-16 (CAT-10-15-19) were based, was as follows:

> Jane Addams cared about people all her life. When she lived in Chicago, she saw many children who did not have a good place to play. They played in the streets. Jane Addams thought the children were not safe. She wanted to do something for them.
>
> Jane Addams was given a large lot. The lot had many old, unused houses on it. She had the houses torn down so that the lot would be empty.

> Jane Addams worked hard to clean up the lot. Then she bought swings and seesaws. This became the first playground in Chicago. The children had a safe place to play!

Five comprehension questions were based on this passage. One of them, and the answer alternatives posed for it, was: What did Jane Addams do with the swings and seesaws?

- cleaned them
- played with them
- had them put on the lot
- used them in the streets

Six out of ten children marked the third answer alternative, which was the intended correct answer. Two marked "played with them" -- apparently reflecting the natural view of children about what one ought to do with swings and seesaws. One child, asked why he chose this answer alternative, explained "Because she could have lots more fun instead of sitting around in the house." One girl marked the first answer alternative. The conversation regarding her answer to the question went as follows:

> L: What about number fifteen?
>
> B: What did Jane Addams do with the swings and seesaws? Played with them, had them put in the lot, used them on the streets. I just said cleaned them.
>
> L: Why did you choose cleaned them?
>
> B: 'Cause she doesn't, she doesn't want the kids to go on dirty things.

According to the publisher's scoring key this is, of course, an incorrect answer. The reading passage does not say that Jane Addams cleaned the swings and seesaws -- but neither does it say that she had them put on the lot. This is not surprising since this item is aimed at testing inferential not literal comprehension. But in terms of plausible inference it seems hard to argue that child B's answer is really less correct than the keyed one. The passage does after all clearly reveal Ms. Addam's concern for cleanliness.

Visual Interpretation. If inferential interpretation was a source of ambiguity in children's perception and reasoning about test items, so was visual interpretation -- specifically how they perceived and reasoned about pictures. Pictures are used in tests for young children far more commonly than in tests for older children. In the tests series included in the pilot study, for example, almost all answer alternatives for science and social studies items (specifically in the CTBS Level C Form S, MAT Primary II Form JS, and SAT Primary Level II Form A) consist of pictures and illustrations among which children are supposed to choose in response to an oral question.* The apparent rationale for such heavy reliance on pictures instead of words is to avoid confounding of science and social studies skills with those of reading.

What the pilot study revealed, however, is that visual interpretation, like reading interpretation, sometimes may be confounded with other skills ostensibly measured by test items. Specifically, different children seem to see different things in the same pictures, and what was seen sometimes influenced how children answer test questions.

---

* The main exception to this general pattern are six sets of items on the MAT for which answer alternatives are posed in terms of the written words "yes" and "no" and "can't tell." We discuss this item type below.

Some of the conversations with children already quoted show that children can perceive or at least describe pictures differently. The relationship between perception and putting names on things is, or course, often unclear. How one perceives something clearly influences the name put on it. The converse may also happen: the names one knows and puts on things may also influence perception. Often, however, the exact relationships between naming and perception are unclear. Thus, when different children identify the same picture as "cabbage," "lettuce" or "leafy vegetable," it is not clear that they are perceiving the picture in fundamentally different ways -- or at least not in ways different enough to lead to ambiguity in terms of answering a test item.

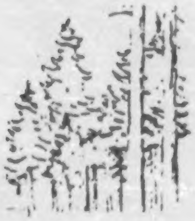
But in some cases perception and naming seems more clearly to influence interpretation. Item 20 from our composite social studies test (CTBS-32-4), provides a relevant example. This item was based on the following passage, written in children's test booklets, but also read aloud to students:

> One of the most powerful rulers of Russia was Peter the Great. During his rule, Peter changed a swampy piece of land into a new capital city and seaport. He also formed the first Russian navy.

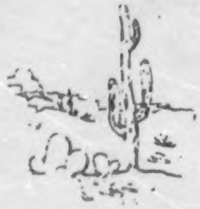
Children were then instructed as follows:

> Find the picture that shows on what kind of land Peter the Great built the capital city. Mark the circle under the picture.

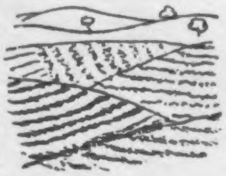
Answer alternatives consisted of the following four pictures. Below each picture are listed the words which children used to describe it, and in parentheses the number of children using each descriptive word or phrase:



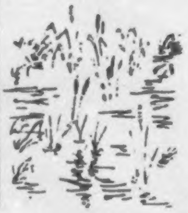
- "forest"(5)
- "woods"(1)
- "some groves"(1)
- "in the woods"(1)
- "like trees, its a forest"(1)



- "desert"(7)
- "a tree"(1)
- "country"(1)



- "mountains" (1)
- "farm"(2)
- "open hills"(1)
- "acres of land"(1)
- "can't tell"(1)
- "country?"(1)
- "nice gardens"(1)
- "like ground"(1)



- "lake"(1)
- "swamp"(5)
- "field...like water sort of"(1)
- "pond"(1)
- "ocean"(1)

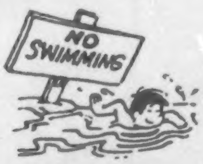
Five of nine children chose the fourth picture which was the keyed answer. Three chose the third picture and one chose the second. It seems significant that three out of four children who got this question "wrong" described the keyed answer alternative not as a swamp, but as field, pond and ocean, and conversely that four out of five children who described the fourth picture as a swamp did choose that answer alternative.

Children's responses to this item also raise questions about what it is that this item reveals about children's understanding and reasoning about social studies. The children who picked the keyed answer may have simply been relying on verbal comprehension -- as one child explained his answer "It said the swamp, and I picked the swamp." However, the children who picked the third alternative seem to have applied broader social knowledge regarding where it is reasonable to build a city. The three children who chose the third answer alternative explained their reasoning as follows:

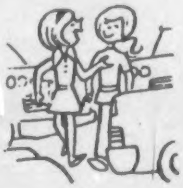
> Child B: "I thought it was right here because it looks like there was more room in that one than any of them."
>
> Child E: "I would say the third one because it has enough room and enough . . . and they could cut the grass and stuff, but it's not much work. 'Cause if it was like in the forest or something, they would have to do all their stuff."
>
> Child F: "I choosed the third one."
> L: "The third one? Why?"
> F: "Because right here, that looks more like a farm to me."
> L: "It looks more like a what?"
> F: "A farm where people could go home."

Differential visual perception does not of course always interfere with children's selection of the intended or keyed answer alternative. This point is well illustrated by item number 10 on our composite social studies test (MAT-21-8). For this item children were orally instructed to "Mark the space under the picture showing someone who is probably following a rule." The three answer alternatives pictured were:

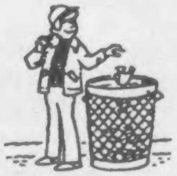
8



○



○



○

All nine children selected the third answer alternative which was the intended correct answer, even though they offered nine different interpretations of the second alternative:

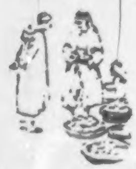
- "two girls that are going on a picnic"
- "girls and they're just dropping things on the ground"

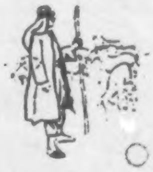
- "two women talking in a parking lot"
- "[two girls] walking through a parking lot"
- "two girls walking between some cars"
- "two ladies walking across a street"
- "I can't tell . . . They're walking in front of a car."
- "Somebody talking to somebody"
- "They are walking when cars are going back and forth. . . "

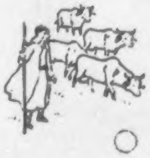
In other items, however, differential visual interpretation clearly affected children's test performance. Probably the clearest example of this occurrence was item 26 on our composite social studies test (CTBS-36-26). Instructions for this item were as follows:

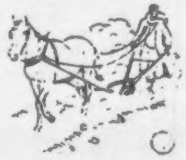
> Look at the story at the bottom of the page. Read the story silently while I read it aloud.
>
> Nomads are people who move from place to place. They make their living by raising sheep and goats. They move to a new place when their herds need more grass and water.
>
> Now look at Number 26. Find the picture that shows how a nomad makes his living. Mark the circle under the picture.

The answer alternative pictures were:


○


○


○


○

Only two children selected the second picture which was the keyed answer alternative. Three chose the third. Three chose the fourth and one child marked both the second and third.

During interviews, two children commented on how hard it was to tell what these illustrations depicted. Comments along these lines were:

> Child A: "Um, I can't see them good."
>
> Child G: "I can't tell in this one, and this one or this one."

In this respect, it should be noted that our practice in constructing the composite tests, namely by xeroxing illustrations from original test booklets, made our illustrations more difficult to interpret than those in the original booklets. Nevertheless, children's descriptions of what they saw in these illustrations clearly indicate that what they saw in answer alternatives influenced which answer they selected.

Two children who chose the third answer alternative. for example, described their choice as follows:

> Child C (who changed his answer during the interview): Now I would say number two, because I just realized those are cows [in number three].
>
> L: Do you remember why you chose number three?
>
> C: Because it looked like somebody watching sheep to me at the time, anyways.
>
> Child G: I think 'cause he, there's goats, and in the story it says goats.
>
> L: Um-hum, and are there goats in that picture?
>
> G: Um-hum, like goats.

The three children who chose the fourth answer alternative explained their reasoning as follows:

> Child A (describing the answer alternatives): Two men and a little girl near the food . . . And a man guarding his sheep. And then a man and the cows are running off, and the horse, a man's on the horse and he's riding going away.

L: OK, which one shows how the
nomads make their living?

A: That one because they keep on
going place to place and place
to place.

Child D (who decided to change his
answer from the fourth to the
third answer alternative, during the interview)

L: OK and which one shows how the
nomad makes his living?

D: The third one.

L: The third one?

D: Yeah.

L: OK, why is that?

D: Because he's raising goats, goats,
goats, goats, goats.

L: Then why did you first choose the
last one?

D: Umm, I didn't know that was goats.

Child H:

L: What's the last picture?

H: Somebody going somewhere.

L: Someone going somewhere. Which one
shows how nomads make their living?
. . . The last one? Why is that?

H: Because, because, em, somebody are
moving somewhere.

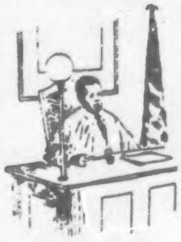
And finally, child I, another child whose answer to this item would have been marked wrong under standard scoring conditions, because he marked both the second and third alternatives, described the answer alternatives as follows:

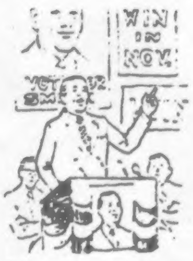
> L: What's going on in each picture?
>
> I: Well, he's buying something from another person, and there he's watching his sheep, and there he's watching his goats. And now he's plowing his lawn or something like that.

In short, in most of these instances children selected ostensibly wrong answer alternatives for the single reason that they saw things in the pictured alternatives other than the test publisher apparently had intended. Some chose the third picture because they saw goats in it and two clearly chose the fourth alternative because they saw this picture as depicting someone "going place to place" and "moving somewhere." These perceptions are particularly interesting because in both cases the children did perceive sheep in the keyed answer alternative. (Child A: "a man guarding his sheep;" Child H: "somebody helping with sheep"), but chose the fourth answer alternative apparently because they perceived moving from place to place as being a more fundamental aspect of how nomads make their living then "guarding" or "helping" with sheep. It is a nice question as to whether such perceptions reveal any less understanding of nomads or social studies than children who picked the keyed answer alternative, perhaps in reaction to the literal statement that "they [nomads] make their living by raising sheep and goats."

There are many other examples of children's descriptions of illustrations which suggest that visual interpretation influenced their perceptions of and answers to test items. Let us recount only one more because it seems to us to be a particularly fascinating example. Item number 24 on our composite social studies test (CTBS-34-24) consisted of the following instructions and answer alternatives:

> 24. Find the picture that shows a person who is running for public office. Mark the circle under the picture.





Only one of nine children selected the fourth answer alternative, which was the keyed or ostensibly correct answer. The eight who got this item "wrong" all picked the first answer alternative. Here is how all nine children described these two answer alternatives, namely the first and fourth picture above:

	First	Fourth
Child A:	"a man in an office"	"president talking"
B:	"a man sitting at a desk"	"he's making a speech"
C:	"looks like somebody in an office"	"somebody giving a lecture"
D:	"a man sitting by a flag at a desk"	"an eye doctor"
E:	"somebody sitting down at a desk or like a principal"	"it's like the president looks, like they're voting for something"
F:	"a guy in his office"	"a man and a judge"
G:	"like a president or a governor, somebody that's a leader"	"like a governor,...a governor talking, speaking, giving a speech"
H:	"somebody looking for public office"	"somebody talking"
I:	"like a judge"	"someone at a concert or something . . . , or talking about a person"

The only child to pick the fourth, keyed, alternative was child A. All the others picked the first; some because they seemed to confuse running for public office, with running an office, but others, such as child H, because they clearly perceived the first answer alternative as somebody looking for public office. Child D who perceived the fourth illustration as an eye doctor instead of someone giving a campaign speech seems to have been clearly led away from the ostensibly correct answer by this perception. Child I explained his selection of the first answer alternative as follows:

> L: The first one? Why is that?
>
> I: 'Cause he, em, a judge, you know, he owns things.
>
> L: He owns things?
>
> I: Well, not really owns things but he owns like big, . . . part . . . I don't know what it's called, but . . . I don't know what it's called . . .
>
> L: OK.

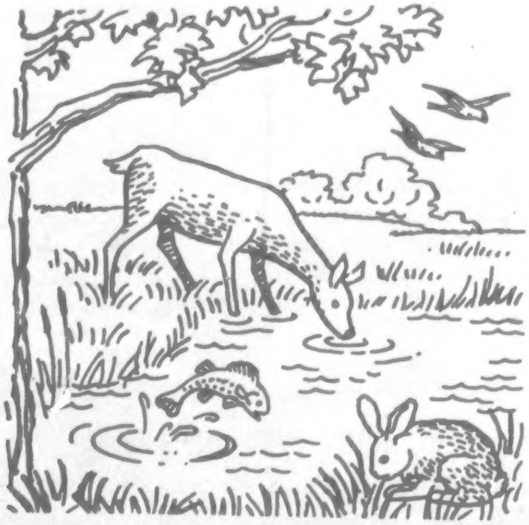
In this case, the child's ability to verbalize and name things clearly seemed to fall short of perception. One can only wish in retrospect that the interview process had been able more clearly to elicit the tie which the child perceived between running for public office and owning things.

Item Format. As discussed already, certain formalities of test-taking like choosing only one answer even if more than one seemed correct, and choosing one even if none seemed appropriate, were apparent sources of ambiguity for some children. In addition one particular type of item format seemed to be a particular source of ambiguity for several children. This type of item, what might be called an information sufficiency format, called for children to listen to a statement while looking at an illustration and then to mark "yes," "no" or " 'can't tell' from the picture what the answer is." In our random sample of items this format appeared in ten items -- six on our composite science test and four on our composite social studies test. All of these items came from the MAT Primary 2, Form JS. This type of item format does not appear in forms and levels of other test series included in the pilot study.

This type of item was the source of six out of eight cases in which children marked more than one answer alternative while taking the test under standard testing conditions. This item type was also one concerning which children frequently decided to change answers during our conversations about these items. To illustrate why this type of item may have been confusing to children let us recount one science item of type.

In item number 17 on our composite science test (MAT-20-28), children were instructed to look at the following illustration and then to mark "yes," "no" or " 'can't tell' from the picture what the answer is" with respect to the following statement:

> The fish had to jump out of the water to get oxygen.



Of eleven children taking the composite social studies test, one marked yes, six marked no, three marked can't tell, and one marked yes, no *and* can't tell, while taking the test under standard conditions. Five children decided to change their answers during our conversations regarding why they answered this item in the way they did. In case the reader is uncertain what the "correct" answer to this question is, we will indicate the publisher's designation of the keyed answer to this item after we recount how the first five children in the pilot study dealt with their own uncertainty on this point.

Child A (who changed his original answer of can't tell to no during the conversation about this item);
L; And number seventeen. The fish nas to jump out of the water to get oxygen.
A; I didn't really hear that 'cause someone was saying something.
L: Oh, would you like to do it now, then?
A: OK.
L: I'll say it again. The fish has to jump out of the water to get oxygen.
A: Nope.
L: No?
A: Not all the time.
L: OK, so you want to choose the answer no?
A: Yes.
L: OK. And why do you think no? How do fish get their oxygen?
A: In the water. They have gills.
L: Um-hum, good. Number . . .
A: But flying fish go up. That's why I wanted to know if it was a flying fish.

Child B:
L: OK, number seventeen. A fish has to jump out of the water to get oxygen.
B; It doesn't (laughs).
L; No, it doesn't. How do you know?
B; If it jumped out of the water to get oxygen then it would probably die.
L; OK.

Child C: I said no to that because fish do not live like we do, you know, they can't get out. They get oxygen under water. Think it's called their gills. I'm not too sure.

Child D: No.
L: Why not?
D: Because fish don't need oxygen. They breathe water.
L: OK
D: I don't know. Some fish can stay out of water, for at least 2 years.
L: Two years - fish can?
D: Some of them.
L: Wow. OK

Child E: (who changed from no to yes during the conversation).
L: In number seventeen, a fish has to jump out of the water to get oxygen.
E: Yeah, that -- ohh.
L: You can change your answer if you want to. What do you want to say on that one?
E: Yes. Because . . .
L: Because you think the fish has to jump out of the water to get oxygen?
E: Well, sometimes.
L: Sometimes? Can you tell from the picture?
E: Uh -- not really.
L: OK. Number seventeen.
E: Seventeen?
L: We filled in seventeen -- you said no -- tell you what, why don't you just put an x through the one you want to choose on that one. Then you don't have to worry about erasing.
E: x in the one . . .
L: In number seventeen. I'll read number seventeen again. The fish has to jump out of the water to get oxygen.
E: I say yes, so I put a x . . .
L: Why not put an x through the yes, 'cause that's your new answer.
E: Yes. There.
L: Why do think the fish has to jump out of the water?
E: Some fish do, not all though, 'cause sometimes fish need oxygen.
L: OK.

The ostensibly correct answer to this question, should there be any doubt in the reader's mind, is no.

Although other yes, no, can't tell items seemed almost as confusing for children as the one cited above, we should hasten to point out that not all items of this type seemed confusing to children. For example, one item asking children to respond to statements concerning numbers on a picture of the earth (e.g. the line marked number 3 is the equator), seemed relatively free of ambiguity.

5.4 The Ambiguity of Ambiguity

We have quoted at some length from transcripts of conversations with children in order to help add meaning to the relatively abstract analyses of test item ambiguity reported in Section IV. In addition, however, these examples have also served to show that there is a real degree of ambiguity in the concept of item ambiguity itself. Our approaches to defining test item ambiguity in terms of the analytical and holistic ratings reported in section IV, were clearly limited in two respects. First, we limited our definition of ambiguity to those instances in which children seemed to evidence the general skill ostensibly measured by an item, but to select an unkeyed answer alternative (unkeyed answer ambiguity) and instances in which children did not evidence the general skill, ostensibly measured, but did still select the ostensibly correct answer (keyed answer ambiguity). In other words, our study has focused on only aspects of ambiguity which were salient enough to affect how children answered test items. There were other aspects of ambiguity revealed in the study (for example, in a reading comprehension passage when all children misperceived the answer alternative "petty" as "pretty," but this did not prevent them from selecting the keyed answer alternative), but these sorts of ambiguity were apparently not prominent enough to affect answer alternatives selected, and hence were not covered in our analyses.

Second in the ratings described in section IV, we relied upon very general statements by publishers of what reading comprehension, science and social studies items were intended to measure (see p. 25-26). Had we attempted to rely upon specific designations of what each item was intended to measure, it is possible that estimates of ambiguity might have been somewhat higher. Recall, for example, the flexible manner in which we interpreted that composite science test item No. 21 was intended to measure either science inquiry skills, or science knowledge (discussed on pp. 61-62).

Clearly there are other ways in which one might define test item ambiguity. For example ambiguity might be defined in terms of inconsistency in children's interpretation of illustrations, reading passages and/or oral instructions. Alternatively one might analyze the degree to which items require or are influenced by children's prior knowledge, when they purport to be measuring something other than such knowledge. In section VI, we will suggest some specific ways in which research on children's perceptions of test items might be extended along these and other lines.

Here, in closing section V, let us simply comment briefly on what are probably the two main sources of ambiguity in the notion of ambiguity explored in the pilot study. Repeatedly we have referred to the phrase test item ambiguity. Yet as we pointed out in section II, the most famous researcher to use the technique of talking with children about test items, namely Jean Piaget, used this technique not to study test items, but to study children's reasoning. What this contrast points out is that ambiguity -- that is doubtfulness or uncertainty of meaning -- resides not in test items per se nor necessarily in children's perception and reasoning per se, but rather in the interaction between the two.

The second general point to be made with respect to the ambiguity of ambiguity is that ambiguity is in essence a relative concept. Apart from death and taxes, little in life is certain. Indeed, if we discount visits from the grim reaper and the IRS man, there may be no social interactions whatsoever which are altogether devoid of doubtfulness or uncertainty of meaning. Specifically, few if any interactions in educational endeavors are altogether free of ambiguity. A good way of illustrating this point is to briefly describe another line of research carried out by Cicourel, MacKay, Mehan, Roth and their colleagues, who as we observed have conducted some of the best research to date on children's interpretations of test items. These investigators (who ironically enough variously describe what they do as sociolinguistics, constitutive ethnography and ethnomethodology) have not only investigated uncertainty of meaning of children's performance on tests, but also have conducted similar investigations of the meaning of classroom lessons. From interviews with teachers and children, and videotape recordings and analyses of classroon interactions they drew the following observations:

> Comparing the teachers' accounts of the lesson before and after it was presented, and comparing the teacher's version with those of the children produced different accounts of the "same" scene. It was sometimes difficult to recognize that the children and the teacher witnessed the same event. The children's responses during the lesson provided different comceptions of correct and incorrect answers which contrasted with the teacher's expectations stated prior to and subsequent to the lesson. The children seemed to receive and organize the lesson in terms of their own orientations at the time of the event, and these conceptions do not always match the teacher's account of the lesson's purpose and conduct.
>
> (Cicourel et al., 1974, p. 4)

So the findings of our pilot study should not be interpreted to mean that ambiguity in children's perception and reasoning about test items is necessarily any greater than uncertainty of meaning in other sorts of educational endeavors. Rather the essential interpretation we would draw is that, as with other interactions in educational settings, and in social life generally, one may see another side to the meaning of test performance by adopting a perspective on it other than the standard one. When test items are viewed from the perspective of children, as we attempted to do in the study reported here, test items do not always seem to represent what is commonly assumed.

The point perhaps needs special emphasis with respect to test items and test scores. Because test items are typically marked right or wrong and seemingly precise numbers are derived from children's performance on them, it is easy to be lulled into a false sense of certainty with respect to their social meaning. But the essential point of our investigation, what might in summary fashion be termed the contextual dependency of children's perception and reasoning about test items, is not a new one. The concern is reflected in even as staid a document as the 1974 Standards on Educational and Psychological Tests, which advises in three "essential" standards regarding test use that:

> - A test score should be interpreted as an estimate of performance under a given set of circumstances. It should not be interpreted as some absolute characteristic of the examinee or as something permanent and generalizable to all other circumstances.
>
> - A test user should consider the total context of testing in interpreting an obtained score before making any decisions (including the decision to accept the score).
>
> - A test user should consider alternative interpretations of a given score.
>
> (APA, AERA & NCME, 1974, standards J1, J1.1 and J7)

On the basis of work reported in this paper, this seems to us like very sound advice. What is needed, perhaps, are better and more widely accepted means of considering the contexts of testing and alternative interpretation of the meaning of test items and scores derived from them.

VI. SUMMARY AND CONCLUSIONS

In this concluding section, we recap what has been said in previous sections, and in part 6.5 offer our conclusions from the pilot study as well as suggestions for research and other activities which might help to clarify some of the issues revealed in the pilot study.

6.1 Background

Critics of standardized tests have charged that test items often are ambiguous or otherwise misleading. Several have offered specific examples to illustrate their criticisms. Defenders of standardized tests have rebutted these criticisms by arguing that 1) item tryout procedures help to ensure that few such items find their way into published tests; 2) items that appear ambiguous or otherwise misleading to adults may not be so to the children for whom they are intended; and 3) even if a few ambiguous or otherwise misleading items do appear in published tests, there is no systematic evidence to show that they constitute a significant proportion of items on any one test or on standardized tests generally.

6.2 The Pilot Study

The pilot study was designed to bring empirical evidence to bear on this dispute. First, we located a group of eleven second- and third-grade children whose parents gave permission for them to participate in the pilot study. Second, we drew a random sample of items from appropriate levels of four of the most commonly used standardized tests (the MAT, CTBS, CAT and SAT). Specifically items were drawn from the reading, science, and social studies subtests of these four test series (except for the CAT which did not have a science or social studies subtest). These subtests were chosen on the theory that ambiguous items would be more likely to show up in these more complex subject areas than in more concrete skill areas like vocabulary and math computations.

Third, these randomly drawn items were put together to make composite reading comprehension, science and social studies tests, and these tests were administered to pilot study children under standard testing conditions. We then talked individually with children about how they had perceived and answered each item. These conversations were audio tape-recorded and subsequently transcribed. Under standard testing conditions, results indicated that on each of the subtests children answered about 70% of the items correctly in terms of the publisher's designated correct answers.

6.3 Analysis

Two types of analyses were performed on the transcriptions of interviews with children -- what we call analytical and holistic ratings. The analytical ratings consisted of having two individuals independently rate each account of how children answered each item using an analytical rating scheme. This scheme dealt with six aspects of children's perception and reasoning about test items: (1) the answer alternative selected; (2) the child's perception of the item stimulus; (3) the child's perception of the response alternatives; (4) the process by which the child appeared to reach an answer to the question; (5) if the child appeared to be reasoning in answering the item question, the apparent validity of the logic used in light of the attribute ostensibly measured by the item; and (6) whether or not the child needed to apply information extraneous to what the item ostensibly measured.

The holistic ratings were organized and carried out somewhat differently. Instead of rating each child's responses regarding each item, under this approach we put together sets consisting of all children's responses regarding one item. We then had independent raters review these item sets, and indicate whether ambiguity was apparent in how one or more children reasoned about each item. Holistic raters were asked to judge four types of potential ambiguity: 1) unkeyed answer ambiguity (occurring when a child marks an "incorrect" answer, but whose description of how he or she answered the question indicates application of the skill ostensibly measured by the item), and keyed answer ambiguity due to 2) merely guessing a keyed or "correct answer"; 3) marking a keyed answer by mistake; or 4) marking a keyed answer for reasons other than mistake or guessing but still not reflecting the skill or attribute ostensibly measured by the item.

Problems of inter-rater consistency were apparent in the initial tryouts of both the analytical and the holistic rating schemes. Some of this inconsistency was, however, due to raters' misunderstanding of particular aspects of each rating scheme. After rating instructions were clarified and ratings corrected for obvious clerical errors, the inter-rater consistency for both rating approaches reached acceptable levels (i.e. inter-rater consistency of 88-90% for analytical ratings and 76-100% for the holistic ratings.*

In both rating schemes, ambiguity in children's interactions with test items was defined as being of two types: 1) keyed answer ambiguity in which the ostensibly correct answer was chosen without demonstration of the skill ostensibly measured and 2) unkeyed answer ambiguity in which an ostensibly incorrect answer alternative was selected even though the skill ostensibly measured was demonstrated in the child's description of how he or she perceived and reasoned about the item. We attempted to distinguish three types of keyed answer ambiguity in both types of ratings -- namely selection of a keyed answer simply by guessing, by mistake or by other reasoning which did not reflect the skill ostensibly measured.

---

*The inter-rater consistencies approaching 100% should not be misinterpreted. They derived strictly from ratings of phenomena which occurred rarely or not at all.

Table 10 summarizes the proportions of types of ambiguity occurring for one or more children in the pilot study as estimated via the two rating approaches. In general, levels of ambiguity identified via the two rating approaches are roughly comparable. The main differences in results from the two rating approaches have to do with the prevalence of unkeyed answer ambiguity. Estimates of proportions of items showing this kind of ambiguity for one or more children were 7% to 18% higher under the holistic rating approach than under the analytic rating approach. The apparent cause of this discrepancy was that while in terms of analytical ratings unkeyed answer ambiguity was defined strictly in terms of four specific rating outcomes, holistic raters appeared to include other phenomena under this rubric, for example cases in which children picked an unkeyed answer by mistake or in conjunction with an unintended perception of item stimuli, but still did seem to demonstrate a valid application of the skill ostensibly measured.

---

[*] It should be noted that for the purposes of presenting results, inconsistencies in ratings were handled differently in the two rating schemes. For analytical ratings, inconsistencies were resolved through discussion between the two analytical raters and one of the principal investigators. For holistic ratings, items were considered ambiguous only if both holistic raters had indicated them to be so for one or more children in the pilot study.

TABLE 10: PROPORTIONS OF ITEMS SHOWING DIFFERENT KINDS OF AMBIGUITY FOR ONE OR MORE CHILDREN ACCORDING TO TWO RATING APPROACHES*

	Keyed Answer Ambiguity			Unkeyed Answer Ambiguity
	Guess	Mistake	Other or Reasoning Invalid With Respect to Attribute	
Analytical Ratings				
Reading	19%	0%	30%	8%
Science	28%	3%	34%	3%
Social Studies	31%	3%	21%	7%
Holistic Ratings				
Reading	11%	0%	30%	22%
Science	31%	0%	24%	21%
Social Studies	41%	7%	38%	14%
Average of Two Rating Approaches				
Reading	15%	0%	30%	15%
Science	30%	2%	29%	12%
Social Studies	36%	5%	30%	10%

---

* Summary data drawn from Tables 7 and 9.

While the latter approach to judging unkeyed answer ambiguity may for these reasons be more appropriate, for the purposes of estimating the generalizability of our findings, we have simply averaged the proportions of ambiguity estimated via both approaches, as shown in Table 10. We then applied statistical tests for confidence intervals around proportions. Since we found confidence intervals for each type of ambiguity to overlap substantially across the three subtests, we pooled estimates across subtests, and estimated confidence intervals across pooled estimates. Using an 80% confidence level, we estimated that had our pilot study sample of children been given the total reading comprehension, science and social studies test forms from which the sample items were drawn, ambiguity of the following forms would have been apparent for one or more children for the following percentages of items:

Keyed Answer Ambiguity	
- Choosing keyed answer by guessing	17 - 37%
- Choosing keyed answer by mistake	0 - 5%
- Choosing keyed answer for other reasons, but without demonstrating skill ostensibly measured	19 - 41%
Unkeyed Answer Ambiguity	6 - 20%

6.4 Discussion

Such findings must be interpreted in light of the fact that our inquiry was only a pilot study and as such was limited in several important respects. Regarding interviews with children, we must point out that simply talking with children may not be a very efficient way of gaining access into their perceptions and reasoning about test items. Also, the particular way we went about interviewing children clearly had its drawbacks. Interviews with children about each item lasted only a minute or two on the average and neither of the interviewers participating in the pilot study had received special training in interviewing children (though both did have several years teaching experience). Nevertheless, even our relatively crude means of inquiry revealed significant levels of ambiguity in children's perceptions and reasoning about test items.

A more severe limitation of the pilot study was the way in which we went about rating our transcripts of conversations with children. In order to provide a basis for generalization, we needed some systematic means of analyzing transcripts of conversations with children. Two rating approaches were developed and tried out. The analytical rating scheme, as explained above, was developed in an effort to analyze each child's account of how he or she reasoned about each item, in terms of several factors which have been suggested to interfere with children's understanding of test items (specifically, how children percieve item stimulus and response alternatives, the process by which they choose an answer, the relevance of reasoning in that process to the attribute ostensibly measured and whether extraneous information not given in the item is required to answer it). The main problem in applying the analytical rating scheme was that in many interviews there was simply not enough information regarding children's perception and reasoning about each item on which to judge each of these issues with much confidence. A second limitation of the analytical rating approach was that we relied on publishers' general statements of what each item type measured rather than specific statements of what attributes were ostensibly measured by each item. Such general statements of attributes intended to be measured clearly caused some confusion for analytical raters.

The holistic rating approach was developed in part to try to overcome some of the problems in the analytical rating approach. Nevertheless, limitations were also apparent in this approach. One limitation, also relevant to the analytical ratings, was the extent to which the interview process itself may have affected children's perceptions and reasoning about items. There were altogether 44 cases in which children decided during interviews to change answers from what they had marked (or not marked) while taking the composite tests under standard conditions (12 cases on the reading comprehension items, 22 on the science and 10 on the social studies). In some cases the changes seemed to be a result of the interview process, in others they seemed spontaneous. What the holistic ratings (and particularly holistic rating inconsistencies) made clear, however, was that it was often difficult to tell when such changes were spontaneous and when they were prompted by the interviewers' questions.

One way of circumventing some of the limitations in the rating schemes employed in the pilot study (although not those in the interview process itself) is to consider directly some examples from our transcripts of interviews with children. Thus in section V, we presented specific examples from transcripts in terms of six apparent sources of ambiguity in children's interactions with test items.

Specifically we provided examples to show how ambiguity (other than that associated strictly with guessing or choosing a keyed answer simply by mistake) seemed to derive from:

- children sometimes forgot the test-taking rule of marking one and only one answer;
- unusual and perceptive interpretation occasionally led children to select an unkeyed answer, even though they seemed to demonstrate skills ostensibly measured;

- personal experience and egocentric interpretation led to keyed answer ambiguity (such interpretation also, of course, often led children to select unkeyed answers, but this would not be included in our definition of ambiguity unless skill ostensibly measured by item was also demonstrated);

- concrete and sequential reasoning caused children to focus on specific details of a problem or a reading passage, and sometimes seemed to lead to both forms of ambiguity (and more generally to cause confounding of assessment of specific skills ostensibly measured with more general reasoning skills, though this is not a point studied directly in the pilot study);

- children often interpreted pictures and illustrations in different ways and this occasionally (though by no means always) led to unkeyed answer ambiguity.*

- one particular kind of item format, namely information sufficiency questions with answer alternatives in the format "yes," "no" or " 'can't tell' from the picture what the answer is" seemed particularly troublesome to some children, and also quite unclear in terms of how the test publisher expected children to interpret such questions.

In most cases it was impossible to say with certainty that ambiguity in children's interaction with test items derived strictly from any one such apparent source. Thus there is a real degree of ambiguity in the notion of ambiguity itself. Specifically, our definition of ambiguity was strictly conditional on whether children picked a keyed or unkeyed answer (without or with demonstration of the skill ostensibly measured, respectively). What the pilot study made clear, however, was that there were other forms and elements of ambiguity not covered under our approach.

---

* One particular limitation of the pilot study should be noted with respect to this finding. The pilot study composite tests were put together from xerox copies of actual test booklets. Since xerox copies of some small scale pictures were less clear than the originals this clearly did make visual interpretation of some pictures more difficult and may therefore have contributed to the diversity of interpretations made of some pictures.

Second, in terms of what items ostensibly measured, we relied upon publishers' general descriptions of what reading comprehension, science and social studies items measured. Thus, a science item might be interpreted to measure science inquiry skills for one child but science knowledge for another. Had we relied upon specific descriptions of what each item ostensibly measured, estimated levels of ambiguity in children's interpretations of test items might have been somewhat higher.

The pilot study clearly shows, nevertheless, that ambiguity does not necessarily reside in particular test items per se nor in children's perception and reasoning per se, but rather in the interaction or lack of congruence between the two. And as a final point of interpretation, it should be noted that ambiguity itself is a relative concept. Doubtfulness or uncertainty of meaning arises not only in children's interactions with test items, but also in children's interactions with teachers, with parents, and certainly in interactions with interviewers, such as those employed in the pilot study.

6.5 Conclusions

The pilot study was designed to bring empirical evidence to bear on disputes concerning the prevalence of ambiguity in standardized test questions. What then can we say in conclusion about this dispute? Keeping in mind that our study was only a pilot inquiry, and was based on only a small, formally unrepresentative sample of second and third graders, we conclude that:

1. Ambiguity in children's interactions with test items is a phenomenon which occurs in a significant number of cases. If defined, as we have

done, as selection of a keyed answer without demonstration of the skill ostensibly measured or selection of an unkeyed answer with demonstration of the skill ostensibly measured, ambiguity was apparent in from 7 - 15% of the total child-item interactions. If an item is defined to be ambiguous when ambiguity of interpretation, as defined above, is apparent for one or more children in the pilot study, then inferential statistics suggest that some 6 - 40% of the items on the test levels and forms included in the pilot study can be estimated to be ambiguous, (if we ignore keyed answer ambiguity caused by children's guessing or marking keyed answers by mistake, see p.58).

2. While ambiguity was apparent in a significant proportion of child-item interactions and a significant proportion of items for one or more children in the pilot study, these proportions were consistently far less than 50%.

In other words, in the dispute over test item ambiguity, as in so many important public controversies, there appears to be some merit to both sides of the argument. Ambiguity appears to be a real phenomenon, but it also appears to be substantially less of a problem than some critics have suggested.

These conclusions may be compared to more traditional means of evaluating test quality, namely reliability and validity. Test reliability traditionally is evaluated in terms of a variety of correlation coefficients or standard errors of measurement. Standard error of measurement refers to the standard deviation of a point estimate of something measured, like a child's standardized test performance. To compare such traditional means of assessing test reliability with results of the pilot study we can calculate .80 confidence intervals (i.e. ± 1.28 x standard error) for each of the test forms included in the pilot study and then as a rough base of comparison to the pilot study, calculate the .80 confidence interval as a percentage of the total number of items on each subtest.

These data are shown in Table 11. These confidence intervals, calculated on the basis of publishers' reported standard errors of measurement for these subtests, amount to 14%-22% of the total number of items on each subtest. From this perspective, the findings of the pilot study appear not too surprising. The analytical rating scheme resulted in estimates of ambiguity as a proportion of total child-item interactions ranging from 7-15% overall (see Table 6). These proportions are well within the levels reported in Table 11 of .80 confidence intervals as percentages of numbers of items on each subtest.*

Nevertheless, issues raised by the pilot study are not essentially ones of reliability or measurement error. Rather they are ones of validity -- namely what it is that test items are measuring for particular children. Specifically the essential point raised in the pilot study is that insofar as we can tell from brief accounts of how children perceived and reasoned about test items, in a significant proportion of cases these items do not appear to be measuring what they are purported to measure. In other words even if levels of ambiguity are within .80 confidence intervals of measurement error, the point is that this is not simply random error. In section V, we discussed some of the extraneous factors which seem to be confounded with children's test performance; namely their adherence to rules of test-taking, unusual and perceptive insights, egocentric thinking in terms of children's own experience, children's attention to concrete detail and sequential reasoning, their differing interpretations of pictures, and their ability to cope with a difficult item format.

Since the pilot study itself carried many ambiguities, we suggest considerable caution in drawing broader conclusions from our study. Nevertheless, we would like to suggest the following implications, with respect to test

---

* It should be noted that we did not apply inferential statistics to the Table 6 estimates since the pilot study sample of children was not a random sample of any broader group.

TABLE 11: NUMBER OF ITEMS, STANDARD ERRORS AND .80 CONFIDENCE INTERVALS FOR SUBTESTS INCLUDED IN PILOT STUDY*

Test Level, Form and Subtest	No. of Items	Standard Error (SE)	.80 Confidence Interval (Z=1.28) (±1.28SE)	.80 Conf. Interval as % of No. of Items
MAT - Primary 2 - JS				
Reading	55	2.9	7.4	14%
Science	40	2.7	7.0	18%
Social Studies	40	2.6	6.6	16%
CTBS - Level C - S				
Reading Comprehension: Passages	18	1.5	3.8	22%
Science	30	2.4	6.2	20%
Social Studies	30	2.4	6.2	20%
CAT - Level 12 - C				
Reading Comprehension	20	1.7	4.4	22%
SAT - Primary II - A				
Reading Comprehension, Part B	48	2.7	7.0	14%
Science	27	2.2	5.6	20%
Social Studies	27	2.7	5.4	20%

---

* Standard error estimates are drawn from publishers' norming data reports for either end of second grade or beginning of third grade, specifically Prescott, et al., Teacher's Manual for Administration and Interpretating, 1978, p.105; CTB/McGraw-Hill CTBS Technical Bulletin No.1, 1974, p.31; CTB/McGraw-Hill CAT Technical Bulletin No.1 (Draft) Table 66; and Madden, R., et al., Technical Data Report, 1975, p.57. Although some publishers report standard errors to the second decimal place, for the sake of consistency all estimates here are rounded to the first decimal place.

development, further research into children's perception and reasoning about tests and test items, and perhaps most important test interpretation and use. With respect to research, we refer briefly to past research relevant to points raised in the pilot study.

Test Development. As a result of our experience in conducting the pilot study and in light of conclusions drawn from it, we completely concur with the recommendations of Anne Anastasi and Edwin Taylor concerning the need for talking individually with children as part of the process of developing and validating tests (see p. 4-5 for their specific recommendations). While traditional means of item tryout and selection (in terms of various item statistics) may help to eliminate some ambiguous or otherwise misleading questions, the results of the pilot study suggest that such procedures fail to screen out significant numbers of items which are ambiguous in that they do not seem to measure what they are purported to measure, for at least some children. Therefore it seems plausible to explore alternative means of test and item validation, including, as Anastasi and Taylor have suggested, talking with children about how they perceive and reason about test questions.

Research.* In addition the pilot study has suggested to us several lines of research which might usefully be pursued. At a minimum, of course, it would be useful to repeat such a systematic analysis of children's accounts of how they perceive and reason about test items in order to test the generalizability of our results. Nevertheless, were we ourselves to undertake another such inquiry, we would alter the study in the following ways. First, we think it would be useful to have children read reading passages aloud. This likely would not help settle questions regarding the component skills of reading comprehension, but it would help clarify, more than we have been able to do in the pilot study, which aspects of reading skills influence children's selection of keyed or unkeyed answer alternatives. Second, as a basis for evaluating the relevance (or validity) of children's accounts with respect to the attribute ostensibly measured, we think in retrospect that it would be better to rely insofar as possible on specific descriptions of what each item is designed to measure rather than on general statements of what particular subtests are designed to measure. In this respect it might be useful to attempt an analysis of what particular items represent for individual children, altogether independent of what items are intended to measure. Results of such an analysis might then be compared on a post-hoc basis with what particular items are intended to measure.

---

* Since we do not attempt to describe research mentioned in any detail, we should point out that almost all of the work cited below is based on types of tests different than the ones treated in the pilot study.

More generally we would like to suggest that these general lines of research might usefully be extended to compare the perceptions and reasonings of children of different age or grade levels. The apparent contrast between our own findings and those of the related investigations of Cicourel, MacKay, Mehan, Roth and Wood (discussed on pp. 9-11) may, for example, be explained by the fact that their studies involved first graders while our own were based on interviews with second and third-graders. Since understanding of and adherence to rules of test-taking seems to be one source of ambiguity in children's perceptions and reasoning about test items, it might be useful to investigate how children, as they progress through school and typically gain experience in taking standardized tests, become socialized into the norms of test-taking. The relevance of such inquiry is clearly supported by the work of Powell which, though highly unusual, indicates that systematic patterns are apparent in wrong answers to standardized test questions by children of different ages and that these patterns exhibit developmental trends (the most recent summary of this work is Powell, 1979. See also Powell, 1968, 1977 and Powell & Isbister, 1974).

Additionally it might prove worthwhile to investigate some of the issues raised in the pilot study in different ways. One could, for example, engage children directly in conversations about what they perceive tests to measure or about what they think of proper ways of taking tests. From their application of this approach with antonym and verbal analogies items from the Scholastic Aptitude Test, Connolly and Wantman (1964) suggested, among other things, that having students talk aloud about their perceptions and reasoning would be a useful means of pretesting such items, and of exploring stylistic differences in students' methods of attacking and analyzing such items.

Also, one could assess the influence of factors such as visual interpretation, personal experience and egocentrism either by employing some independent background information on children interviewed or by manipulation of different aspects of test items. For example, if children's performance on our reading comprehension item No. 1 (Eve likes to watch TV. She thinks it is (blank). Answers: fun, dull, hard, petty), is influenced by the fact that most children think watching TV is fun, it would be interesting to compare performance on the item as given with an item which read "Eve hates to watch TV. She thinks it is (blank)."

A variety of past investigations suggest the value of such manipulation as a means of investigating children's perceptions and reasoning about tests. Bally's (1976) study with second grade children, for example, showed that a standardized test of auditory discrimination confounds measurement of the skill ostensibly measured with the familiarity of words used to test that skill. Similarly, Franklin's (1974) study with first and second grade children suggests that picture preference may be a significantly influencing factor on word picture vocabulary tests typical of achievement and aptitude tests used in the early grades. Also, a variety of passage dependency studies, in which children were asked to answer reading comprehension test items without reading the passages on which they were based, have raised questions about the extent to which such items measure attributes other than reading comprehension. (Of more than a dozen references on this topic, Scherich & Hanna, 1977; and Hanna & Oaster, 1978, 1978-79, are three of the more recent.)

Test Interpretation and Use. While new methods of research and test development, including talking with children about test items, may hold some potential for improving the quality of test items, it seems to us that this potential may be limited by two factors, the first having to do with how test quality is judged, and the second with the sometimes idiosyncratic nature of children's perception and reasoning about tests.

Regarding standards for judging test and test item quality, let us refer to Thorndike's Educational Measurement (1971) which is probably one of the most authoritative sources of advice on a variety of testing issues. In a chapter giving advice on writing test items, Wesman (1971) suggests "a portion of the art of item writing is the ability to conceive of distractors which are incorrect by any standard, and yet are attractive to those not knowing the answer" (p. 82). In a later chapter discussing pretryouts of test items, Henrysson (1971) suggests that if an item distractor--that is an unintended response alternative--"is obviously a wrong answer to almost everyone, and thus very implausible", then "this distractor probably should be replaced by a more plausible answer" (pp. 136-7). Obviously there is a real tension implicit in such advice. How is it that an unintended answer alternative or distractor can be "wrong by any standard" but still be attractive or plausible to those not knowing the intended answer? The point of this contrast is that it represents a real ambiguity in the norms of standardized test construction. Typically, in the process of test development less priority is given to direct evaluation of what items reveal about what test-takers do or do not know, than to indirect evaluations in terms of item discrimination indices regarding either whole test scores or external criteria. The test-score indices corresponding to these item statistics are coefficients of internal consistency and criterion-related validity. It is just such indices in terms of which the technical quality of tests typically is judged. Thus, it seems to us that if talking with children about how they perceive and reason about test items does have potential for improving test quality, it will not likely be so much in terms of improving such traditional indices of test quality, as in clarifying the meaning of test items, both for children who take them and adults who try to view and understand children through them.

The second limitation regarding the potential value of interviewing children as a means of test improvement is simply the diversity of how items are perceived and attacked. The pilot study has, we think, raised some important general issues about children's concrete reasoning and egocentrism in dealing with test items, but some of the interpretations drawn by children in the pilot study, particularly those reflecting unusual personal experience, probably would be quite rare. In response to the science question about where a heavy boy and a lighter boy should sit to balance a seesaw, for example, how often might it happen that children would explain the choice of the unkeyed answer alternative showing boys sitting at far ends of the seesaw (see p. 71) in quite the same way as Child B:

> Well, usually when my little brother and I get on the seesaw we sit at the ends because if I sit closer then we might -- he'll say, "get back, get back." He's afraid he'll fall.

But if unique and idiosyncratic interpretations of test questions by individual children would make the task of test-developer in trying to eliminate ambiguity seem never-ending, they would seem to hold considerable potential, simply by way of illustration, for helping to prevent mis- or overinterpretation of test scores. As Judge Skelly Wright remarked in the Hobson v. Hanson case in Washington, D.C. in 1967, "Although test publishers and school administrators may exhort against taking test scores at face value, the magic of number is strong" (Hobson v. Hanson 269, F. Supp. at 489). Testing experts often attempt to prevent misplaced faith in the precision of test scores by explaining statistical concepts such as standard error of measurement and test score reliability. It seems to us that concrete illustrations of the unique ways in which children may interpret test questions may be a more effective means of communicating such concerns to those unfamiliar or uncomfortable with statistics.

In this respect, we would like to suggest too that it might be quite constructive, not only for test publishers, but also for those who regularly engage in administering or using test scores, to try talking with individual test-takers about how they perceive and interpret test scores. Teachers who have used this approach, such as Ann Cook (Cook & Meier, n.d.) attest to the value of this approach, and there is at least some research evidence (Shemer, 1975) to suggest that teachers rate information on how and when children understand test questions to be more useful than traditional test scores. In short talking with children about the meaning of tests not only might help people to avoid succumbing overmuch to the magic of numbers and test scores, but also would, we suspect, help more generally to remind them, as the pilot study has reminded us, of how different the worlds of children and adults can be.

REFERENCES

American Psychological Association, American Educational Research Association and National Council on Measurement in Education. Standards for educational and psychological tests. Washington, D.C.: American Psychological Association, 1974.

Anastasi, Anne. Principles of psychological testing. New York: Collier Macmillan Publishers, 1976.

Bally, E.B. The effects of levels of word familiarity on tests of auditory discrimination and its relationship with socioeconomic level and race. Unpublished doctoral dissertation, Catholic University of America, 1976. (University Microfilms No. 76-19, 362)

Block, N., & Dworkin, G. (Eds.). The IQ controversy. New York: Pantheon, 1976.

Cicourel, A., et al. Language use and school performance. New York: Academic Press, 1974.

Connolly, J., & Wantman, M. An exploration of oral reasoning processes in responding to objective test items. Journal of Educational Measurement, June 1964, 1(1), 59-64.

Cook, Ann, & Meier, Deborah. Reading tests: Do they help or hurt your child? New York: Community Resources Institute, n.d.

CTB/McGraw Hill. Comprehensive tests of basic skills (expanded edition). Monterey, Ca.: Author, 1973, 1974, 1976, 1977.

CTB/McGraw Hill. California achievement tests. Monterey, Ca.: Author, 1977.

Dimengo, Carl. Basic testing programs used in major school systems throughout the United States in the school year 1977-78. Akron, Ohio: Akron Public Schools Division of Personnel and Administration, March 1978.

Dunnelle, M.D. Critics of psychological tests: Basic assumptions: How good? Psychology in the Schools, 1964, 1, 63-69.

Ebel, R.L. What do educational tests test? Educational Psychologist, 1973, 10, 76-79.

Finley, C., & Berdie, F. The national assessment approach to exercise development. Ann Arbor, Mi.: National Assessment of Educational Programs, 1970. (ED 067 402)

Franklin, E.R. The effect of a picture on a child's error on a test. Unpublished doctoral dissertation, Columbia University, 1974. (University Microfilms No. 74-28, 493)

Freund, J. Modern elementary statistics (4th edition). Englewood Cliffs, N.J.: Prentice-Hall, 1973.

Gamel, N. et al. State ESEA Title I reports: Review and analysis of past reports and development of a model reporting system and format. Mountain View, Ca.: RMC Corporation, October 1975.

Gruber, H., & Voneche, J. The essential Piaget. New York: Basic Books, 1977.

Hanna, Gerald S., & Oaster, Thomas R. How important is passage-dependence in reading comprehension? Journal of Educational Research, July/August 1978, 71, 345-348.

Hanna, Gerald S., & Oaster, Thomas R. Toward a unified theory of context dependence. Reading Research Quarterly, 1978-1979, XIV(2), 226-243.

Henrysson, S. Gathering, analyzing and using data on test items. In R. Thorndike (Ed.), Educational measurement (2nd edition). Washington, D.C.: American Council on Education, 1971.

Heyneman, S., & Mintz, F. The frequency and quality of measures utilized in federally sponsored research on children and adolescents. Washington, D.C.: Social Research Group, George Washington University, 1976.

Hobson v. Hansen. 269. Federal Supplement 401 (D.D.C. 1967).

Hoffman, Banesh. The tyranny of testing. New York: Crowell-Collier Press, 1962.

Houts, Paul L. (Ed.). The myth of measurability. New York: Hart Publishing Company, 1977.

Jensen, A. Bias in mental testing. New York: Free Press, 1980.

Krejcie, R., & Morgan, D. Determining sample sizes for research activities. Educational and Psychological Measurement, 1970, 30, 607-610.

Lennon, R. Perspective on intelligence testing. Measurement in Education, Spring 1978, 9(2), 1-8.

Luchins, A.S., & Luchins, E.H. Wertheimer's seminars revisited: Diagnostic testing for understanding of structure. In P.W. Johnson-Laird and P.C. Wason (Eds.), Thinking readings in cognitive science. New York: Cambridge University Press, 1977.

MacKay, Robert. Conceptions of children and models of socialization. In Hans Peter Dreitzel, Childhood and socialization. New York: Macmillan Publishing Company, Inc., 1973.

MacKay, Robert. Standardized tests: Objective/objectified measures of competence. In Aaron Cicourel, et al. Language use and school performance. New York: Academic Press, Inc., 1974.

Madden, R., Gardner, E.F., Rudman, H.C., Karlsen, B., & Merwin, J. Stanford achievement test. New York: HBJ, Inc., 1972, 1973, 1975.

Matarazzo, J. Wechsler's measurement and appraisal of adult intelligence. Baltimore: Williams & Wilkins, 1972.

Mehan, H. Assessing children's language using abilities. In J. Armer & A. Grinshaw (Eds.), Methodological issues in comparative sociological research. New York: Wiley, 1973.

Mehan, H. Structuring school structure. Harvard Educational Review, 1978, 48(1), 32-64.

Mehan, H., & Wood, H. The reality of ethnomethodology. New York: Wiley-Interscience, 1975.

Meier, Deborah. What's wrong with reading tests? Workshop Center for Open Education, March 1972.

Nisbett, R.F., & Wilson, T.D. Telling more than we can know: Verbal reports on mental processes. Psychological Review, May 1977, 84(3), 231-259.

Patton, M. Understanding the gobble-dy-gook: A people's guide to standardized testing. In V. Perrone & M. Cohen (Eds.), Testing and evaluation: New views. Washington, D.C.: Association for Childhood Education International, 1975.

Powell, J.C. The interpretation of wrong answers from a multiple choice test. Educational and Psychological Measurement, 1968, 28(2), 403-412.

Powell, J.C. The developmental sequence of cognition as revealed by wrong answers. Alberta Journal of Educational Research, 1977, 23(1), 43-51.

Powell, J.C. Can developmental status information be obtained from wrong answers? Paper presented to the Annual Conference of the National Council on Measurement in Education, San Francisco, April 1979.

Powell, J.C., & Isbister. A comparison between right and wrong answers on a multiple choice test. Educational and Psychological Measurement, 1974, 34(3), 499-509.

Prescott, G.A., Balow, I.H., Hogan, T.P., & Farr, R.C. Metropolitan achievement tests. New York: HBJ, Inc., 1977, 1978.

Roth, D. Intelligence testing as a social activity. In A. Cicourel et al., Language use and school performance. New York: Academic Press, 1974.

Scherich, Henry H., & Hanna, Gerald S. Passage-dependence data in the selection of reading comprehension test items. Educational and Psychological Measurement, 1977, 37, 991-997.

Schwartz, J. A is to B as C is to anything at all: The illogic of IQ tests. In P. Houts (Ed.), The myth of measurability. New York: Hart Publishing, 1977.

Shemer, R. Testing the limits with the Stanford-Binet intelligence scale for children. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, 1975. (University Microfilms No. 76-6956)

Snedecor, G., & Cochran W. Statistical methods. Ames, Iowa: University of Iowa Press, 1972.

Taylor, Edwin F. The looking glass world of testing. In Standardized testing issues: Teachers' perspectives. Washington, D.C.: National Education Association, 1977a.

Taylor, E. "Science Tests" 1977b. In Houts, P. The Myth of Measurability. New York: Hart, 1977b. pp 291-308.

Wesman, A. Writing the test item. In R. Thorndike, Educational measurement (2nd edition). Washington, D. C.: American Council on Education, 1971.